frontiers | Research Topics

Understanding in the human and the machine

Edited by
Yan Mark Yufik, Karl Friston and Rosalyn J. Moran

Published in
Frontiers in Systems Neuroscience

ISSN 1664-8714
ISBN 978-2-88976-823-3
DOI 10.3389/978-2-88976-823-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Understanding in the human and the machine

Topic editors
Yan Mark Yufik — Virtual Structures Research Inc., United States
Karl Friston — University College London, United Kingdom
Rosalyn J. Moran — King's College London, United Kingdom

Cover image
Fedorov Oleksiy/Shutterstock.com

Citation
Yufik, Y. M., Friston, K., Moran, R. J., eds. (2023). *Understanding in the human and the machine*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-823-3

Table of contents

frontiers | Frontiers in Systems Neuroscience

TYPE Editorial
PUBLISHED 25 November 2022
DOI 10.3389/fnsys.2022.1081112

OPEN ACCESS

EDITED AND REVIEWED BY
Heiko J. Luhmann,
Johannes Gutenberg University
Mainz, Germany

*CORRESPONDENCE
Yan M. Yufik
imc.yufik@att.net

RECEIVED 26 October 2022
ACCEPTED 02 November 2022
PUBLISHED 25 November 2022

CITATION
Yufik YM, Friston KJ and Moran RJ (2022) Editorial: Understanding in the human and the machine.
Front. Syst. Neurosci. 16:1081112.
doi: 10.3389/fnsys.2022.1081112

Editorial: Understanding in the human and the machine

Yan M. Yufik[1]*, Karl J. Friston[2] and Rosalyn J. Moran[3]

[1]Virtual Structures Research Inc., Potomac, MD, United States, [2]Queen Square Institute of Neurology, University College London, London, United Kingdom, [3]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

KEYWORDS

self-organization, understanding, grasp, general intelligence, complexity, prediction, explanation, mental model

Editorial on the Research Topic
Understanding in the human and the machine

This Research Topic was initiated in a workshop—in August 2021 in Washington D.C. – under the auspices of the U.S. Air Force Office of Scientific Research and Air Force Research Laboratories. This Issue is dedicated to analyzing understanding and is a sequel to 2017 Research Topic, which focused on the fundaments of self-organization in the nervous system https://www.frontiersin.org/research-topics/4050/self-organization-in-the-nervous-system#articles. A crosscutting theme in both journals—and the workshop—is the principle of Variational Free-Energy Minimization (VFEM), also known as Active Inference (Friston et al., 2006; Friston, 2010; Parr et al., 2022). This principle has been applied to further our understanding of the role, adaptive value and neuronal mechanisms of the capacity to understand ("understanding of understanding"). Conceptualizing understanding as a product of uniquely human self-organization—obtaining levels of free energy minimization inaccessible to other species—appears to offer a promising perspective on the neuronal underpinnings of understanding and designing devices possessing a modicum of human understanding (machine understanding). This editorial reviews the state-of-affairs in the multidisciplinary domain of understanding R&D ("the science of understanding"), summarizes some key ideas in theoretical approaches centered on the application of VFEM (Yufik and Friston, 2016; Yufik et al., 2017), and introduces contributions in the present collection.

Human intellect apprehends the world and itself through the lens of understanding. Since the time of Aristotle (2004) the capacity to understand—and the innate desire to exercise that capacity—have been recognized as the defining features of human intelligence, distinguishing humans from other species (Lear, 1988; Greco, 2014). Analysis of how understanding operates and influences the ways humans interact with the world—and with each other—has remained a key focus in psychology (Piaget, 1974, 1978) and philosophical discourse throughout history [e.g., (Kant, 1990 (1781); Hegel, [1977 (1807)]; Locke, [1996 (1689)]; Russell, [1997 (1921)]; Descartes, [1998 (1637)]; Berkeley, [1998 (1734)]; Hume, [2018 (1739)]). Although never dormant, interest in the phenomenon of understanding was renewed and re-invigorated in the modern era, due to the emergence of radically novel conceptual constructs in mathematics, physics, biology, psychology and other disciplines turning to "eternal"

questions like what makes the world understandable, the origins and limits of understanding, etc. from the realm of speculative philosophy to the mainstream of scientific inquiry (Mehra, 1999; Freeman, 2000; Barsalou, 2008; Rovelli, 2014; de Regt, 2017). Accomplishments in the last decades—at the intersection of computer science, neuroscience and other disciplines—have realized some intelligence (learning, reasoning) in engineering artifacts. The resulting proliferation of smart systems, including weapons capable of acting autonomously or collaboratively with warfighters, has created an urgent demand for advances in machine intelligence to furnish a competitive edge in commerce and defense. This Research Topic seeks to facilitate progress in the science of understanding, with a special focus on machine understanding.

What is understanding and how does it effect performance? Continuing debates on the subject (Gelepithis, 1986; Baumberger et al., 2016; Hannon, 2021) reveal a tangle of issues and controversies that can be traced back to Plato and Aristotle. And have not been settled since. In particular, difficulties persist in clarifying relations between understanding, knowledge and belief (Grimm, 2006; Baumberger, 2014; Pritchard, 2014), defining the value (benefits) of understanding in adaptive performance (Kvanvig, 2003, 2009; Grimm, 2012, 2014), circumscribing the relative roles of explanation and prediction enabled (and perhaps entailed) by understanding (Khalifa, 2013). The cognitivist school in psychology reduces understanding to possessing algorithms (subject S understands task T if S possesses algorithms for carrying out T) (Newell and Simon, 1972; Simon, 1979). Conversely, other authoritative sources maintain that understanding involves non-algorithmic and non-computable components (Penrose, 1997, 2016) and argue that algorithms can be designed so that computers give the impression of understanding a task, while remaining clueless about its meaning (Searle, 1990; Kauffman, 2010). An example from a psychology classic (Piaget, 1978) illustrates the distinction between the way non-algorithmic and algorithmic processes manifest: consider a row of N domino pieces standing on edge and compare two kinds of performance: predicting at a glance that, whatever N, when pushing the first piece, the last piece will fall, vs. predicting the same but only after having worked mentally through all the N pieces, one-at-a-time. According to our proposal, diverging views on understanding are not mutually exclusive but reflect different components and operational stages in the underlying mechanism, as discussed below.

Variational Free Energy Minimization (VFEM) rests on several assumptions, including the following: (a) to survive, any organism, from the simplest (bacteria) to most advanced (humans), must possess internal (*a.k.a.*, world or generative) models that embody regularities in the organism's environment, (b) such internal models stir an organisms' interaction with the environment toward minimizing variational free energy (VFE) in sensing–acting cycles (roughly speaking, the VFE expresses prediction errors, that is, discrepancies between sensations predicted to follow actions and those actually experienced) and (c) suppression of prediction errors goes hand-in-hand with resisting entropic forces and maintaining organisms in characteristic states (of low entropy) (Friston, 2010). Our contention is that understanding engages particularly efficient mechanisms that are unique to human brains. Interested readers can find more detailed discussions of these notions in Yufik and Friston (2016) and Yufik (2019, 2021a,b). In brief:

To appreciate the distinction between understanding and learning, consider how different approaches account for superior performance in chess. The learning-centric approach attributes such performance to assimilating large stores of chess data and winning a new game with reference to the winning moves of previous games (Chase et al., 1973; Gobet and Simon, 1996). This account leaves unexplained how humans can compete with machines that have access to unlimited data and operate with processing rates billions of times faster than those seen in humans. Particularly mystifying is a quite common phenomenon of young talent defeating adult masters [e.g., a 9 year old Reshevsky played over 1,500 games of simultaneous chess in one US tour and lost <0.5% of the games (Reshevsky, 2012)]. An alternative view predicates superior performance on superior understanding. How so?

Three definitions in the literature identify significant components of the understanding capacity (with some critical exceptions, as will be explained shortly):

1. "Understanding, grasp: apprehending general relations in a multitude of particulars" (The Webster's Collegiate Dictionary).
2. "Understanding requires the grasping of explanatory and other coherence-making relationships in a large body of information. One can know many unrelated pieces of information, but understanding is achieved only when informational items are pieced together" (Kvanvig, 2003, p. 192).
3. Scientific understanding involves expressing relations in the form of equations and acquiring "some feel for the character of the solution …. if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we "understand" the equation, as applied to the circumstances" (Richard Feynman, cf. de Regt, 2017, p. 102)

A simple example serves to illustrate these definitions. Consider a scene comprising just two "particulars" (dog, cat) and imagine grasping the relation between them: "dog *chasing* cat." Note that such grasping requires (a) recognizing individual behaviors (running cat, running dog), (b) piecing these informational items together (Kvanvig, 2003, p. 192) and (c) apprehending a particular form of behavior coordination (*chase*). Grasping the relation brought about "a way of knowing

what should happen in given circumstances" (Richard Feynman, c/f de Regt, 2017, p. 102) which includes prediction (e.g., if the dog runs faster than the cat it will intercept the cat; if the cat speeds up, so will the dog, etc.) and explanation (e.g., the cat is running because it's being chased by the dog). Such rough (qualitative) predictions are inherent in—and derive directly from—the relation, and can be followed by reasoning about details, in order to achieve better prediction accuracy (e.g., "solving equations" to determine the time of intercept given the distance and velocities).

Consider now increasing the number of "particulars" and complicating the scene in three ways: (a) imagine that the cat disappears behind a fence, (b) let there be an observer trying to predict what might happen and let there be a tree behind the fence, visible to the observer and (c) imagine the observer seeing no trees but entertains the possibility of their presence. In (a), the dog changes course and runs to the other side of the fence to intercept the cat. In (b) and (c), the dog's behavior does not change, but the observer realizes that the cat might climb the tree and thus leave the dog disappointed. Predators are genetically equipped with modeling mechanisms that reflect long-term statistical averages in the behavior of their prey (e.g., on the average, prey continue their movement patterns when disappearing behind objects) and allow gradual response tuning in the vicinity of such averages, based on individual experiences (learning). Such mechanisms restrict adaptive behavior to recollecting precedents—if available—or to trial-and-error, otherwise (i.e., error suppression strategies in (b) and (c) are not accessible to most creatures). By contrast, human mechanisms support the composition of unified relational structures that integrate the recollected, and current sensory elements, and simulate interdependencies among them. Understanding overcomes restrictions engendered by both genetically fixed automatisms and individual learning—and makes possible predicting and constructing adequate responses to novel conditions. The mechanism engages three key components (Yufik, 1998, 2013, 2021a,b):

1. Integration of initially unrelated elements into coherent relational models in one-step transitions (akin to phase transition in physical substrate),
2. Models are synergistic structures: they impose coordination between the constituent elements that constrain their variation,
3. Models are self-coordinating and resist fragmentation.

Some clarifications are called for here.

1. Borrowing the notion from physics, models can be viewed as *virtual systems* (Yufik, 1998) holding a superposition of possible organizations afforded by the arrangement of elements (e.g., and expert model of piece arrangements on a chessboard holds a superposition of plausible piece grouping (or functional complexes, in the sense of De Groot) (De Groot, 1965). Such superpositions collapse to one configuration yielding the steepest entropy reduction in the virtual system, giving rise to the experience of *grasp*, e.g. [(cat running somewhere), (dog running somewhere)] → (dog *chasing* cat)].
2. Collapse and compression establish coordination across the model that suppresses superfluous (redundant) variations. For example, a thought that the cat might start grooming does not cohere with the form of behavioral coordination determined by the relation, which bars such thoughts from entering the observer's mind when predicting outcomes.
3. In unified models, thinking of variations in one element effects corresponding variations in others (hence, the self-coordination). For example, envisioning the cat climbing the tree immediately implies a failure to intercept. Similarly, when considering the moves of particular pieces, unified models—held by experts—render them aware of the accompanying exposure and changing relations across the board, while fragmented models (c.f., novices) preclude such awareness (Yufik and Yufik, 2018). To intuit the difference, think of taking opponent's piece and loosing the game in a few moves ("fool's mate") vs. sacrificing own piece and winning the game.

Crucially, compression and self-coordination in models precludes an inefficient wasting of time and energy on (considering) actions with marginal or no impact, while keeping in focus those few that decide the outcomes—the actions that "matter." The scale of such savings can become astronomical as the number of elements increases. Studies of expert performance in complex dynamic tasks (firefighters, military commanders) have found that expert decision processes, instead of weighing alternatives, converge quickly on a single plan considered by them to be "obvious" (Klein, 2017). In a similar vein, possibilities and risks inherent in piece arrangements can be obvious to a chess prodigy, while less capable players are forced to move step-by-step through combinatorial fog. A lack of understanding turns chess positions into incoherent arrangements of pieces, each having several degrees of freedom. In contrast, expert models "squeeze out" degrees of freedom and thus provide "a way of knowing what should happen in given circumstances" (Richard Feynman, c/f de Regt, 2017, p. 102).

Summarily, understanding derives from self-organization in the brain that amplify adaptive efficiency, by supporting the construction of models representing objects, their behavior and patterns of behavioral coordination—and enabling an increase in the expressive complexity of such models, without compromising their efficient use. Stated differently, human models enable prediction and construction of apt responses to complex interplays between multiple environmental entities, by collapsing combinatorial spaces engendered by those

interplays Complexity collapse (radical simplification) makes complex situations and responses to them meaningful and explainable (Yufik, 1998, 2002, 2013). Activities in neuronal masses constitutive of such models remain the subject of current and future research (Moran et al., 2013). This kind of efficiency emerges in the minimization of VFE *via* the implicit maximization of model evidence or marginal likelihood associated to the internal model. In this formulation, log model evidence can be expressed as accuracy minus complexity. This means, minimizing VFE is simply a description of the kind of sentient behavior considered above; namely providing an accurate account of exchange with the world that is as simple as possible. Understanding is the key to the right kind of complexity minimization—the right kind of collapse across degrees of freedom that capture the regularities, invariances and compositional regularities evinced by our [inter]action with the lived world. Indeed the aging brain may imbue better understanding through increased generalizability (decreased complexity, Moran et al., 2014).

We now turn to the contributions in this Research Topic. While centered on VFEM formulations, the intent for the Issue was to showcase current thinking about understanding and related problems. Accordingly, articles in the Issue address a range of opinions spanning philosophy, neuroscience, cognitive science, biology and engineering, with an excursion into biological underpinnings of cognitive pathologies. This introduction serves as an annotated table of contents, breaking the collection into several (overlapping) thematic groups.

Philosophy of understanding

Khalifa et al. discuss the relative roles of philosophy and other disciplines (cognitive science, neuroscience, other) in advancing the science of understanding, suggesting that philosophy can offer a framework for both formulating discipline-specific accounts of understanding and then unifying such accounts under a general theory. Sloman et al. argue that inquiry into biological foundations of human intelligence should not be confined to analyzing individual brains but must consider communities of individuals.

Understanding and consciousness

Pepperell considers whether progress in machine understanding is predicated on advances machine consciousness, leaning toward answering this question in the affirmative. Arguments encompass both general ideas and experimental findings in neuroscience, venturing into the domains of creative thinking (understanding paintings) and offering suggestions regarding the limitations of machine learning and requirements for machine understanding. Luczak and Kubo examine the relations between consciousness and adaptive efficiency. Their predictive Neuronal Adaptation hypothesis associates consciousness with prediction and ascribes prediction and error correction abilities to individual neurons—acting as basic functional units—that underwrite consciousness.

Human-machine interaction

Parr and Pezzulo observe that applications of machine intelligence are hampered by the machine's inability to explain its decisions, and engage VEFM to argue that comprehensive explanations require the optimization of generative models at two levels: a model of the world chooses responses based on the predicted conditions in the world and a higher-level model predicts choices in the world model and uses such predictions to formulate explanations of the lower-level decisions. Schoeller et al. observe that the robustness of human-machine interaction depends on the level of trust experienced by users, and analyze trust determinants and trust-building strategies from the vantage point of VFEM. Blaha et al. point at the existence of different stages in the process of reaching understanding, and suggest natural language probes for tracing progress through the stages expected to be conserved over humans, machines and human-machine teams. Llinas and Malhotra review current research on situation control and suggest approaches, in the spirit of the VFEM, toward expanding research scope, focusing on the construction of adaptive situation models that can predict situational changes and then use prediction outcomes to minimize errors. Yufik and Malhotra. discuss distinctions and overlap in the notions of "situation awareness" and "situation understanding" and argue that attaining mutual human-machine understanding requires establishing an isomorphism between the corresponding models. More precisely, since human models represent objects, their behavior and forms of situated behavioral coordination, machine models that represent the same would be inherently explainable to users and would allow straightforward mapping of user feedback onto machine processes (hence, the mutual understanding).

Evolutionary origins

Vicencio-Jimenez et al. discuss the thermodynamic aspects of cognitive processes and propose Energy Homeostasis Principle (EHP) complementing the VFEM principle in explaining the origins and evolution of intelligence. Intelligence develops in an open thermodynamic system (brain) in a growing hierarchy of components (neuronal groupings) that regulate their energy needs and interact with other components in the hierarchy while preserving a degree of independence. Kozma et al. rely on a vast amount of EEG data to formulate

a model of neuronal processes underlying intelligence. EEG recordings demonstrate self-organization of neuronal activities, interspersed with episodic collapses in the ensuing structures. Such local phase transitions produce phase gradients that correlate with transient perceptual experiences. The mechanism of phase transition and become being gradient propagation is consistent with those envisioned in the Global Workspace Theory and may be responsible for optimizing trade-offs between demands posed by rapid adaption to novelty vs. preservation of stability. Latash discusses substantive similarities in the theories of motor control and cognitive control: both postulate predictive processes and anticipatory adjustments to actions and assume that such prediction and adjustments are carried out by self-organization processes in the control system, particularly producing task-specific synergistic groupings of control elements. These similarities may be indicative of a common synergistic mechanism participating in the entire range of control activities, "from figuring out the best next move in a chess position to activating motor units appropriate for implementing that move on the chess board" (Latash, this Research Topic).

Cognitive architecture

Kroger and Kim investigate neuronal responses in frontopolar cortex (FPC) known to participate in the performance of complex cognitive functions, including understanding. The study seeks to determine differences in FPC involvement when subjects respond to two types of demands: acquiring and maintaining structured information vs. manipulating such information in performing cognitive tasks. Analysis of fMRI data reveals differences in FPC recruitment and activities sensitive to task organization and complexity. FPC appears to be particularly involved when responding to new and/or creating new information. Safron et al. describe a bio-inspired architecture for robotic control. Analysis of cognitive control focuses on the navigation problem involving simultaneous localization and mapping (SLAM) (i.e., build a map of the terrain concurrently with identifying one's location on the map) and hypothesizes that navigation mechanisms residing in the hippocampal/entorhinal system could be coopted by evolution in the implementation of higher cognitive functions. Construction of the world model in the SLAM architecture is governed by the VFEM principle, entailing optimization of representational units (c.f., categories) in the model.

Machine learning

Articles in this thematic group illustrate application of machine learning methods in the type of tasks where they excel the most, i.e., classification and recognition. Cai et al. review results in the application of machine learning and feature extraction algorithms for emotion recognition, that is, classifying EEG signals and correlating such classes with emotional states of the subjects, following classifications of discrete emotional states in psychological literature. Wang and Zeng use learning in Spiking Neural Networks (SNN) to model acquisition of concepts integrating features of different sensory modalities (multisensory concept learning), under two conditions: preceding integration, inputs in each modality either become associated, or remain independent. Integration vectors produced by the SSN procedure are subsequently labeled (correlated to concepts) by psychologists.

Neurobiological mechanisms of cognitive pathologies

Wang et al. investigate pathological conditions in the nervous system of schizophrenia patients that cause grossly maladaptive behavior (severe aggression) and admit correction only *via* medical treatment. Having established the correlation between aggression severity and inflammation accompanied by bacterial dislocation, the study suggests development of novel methods for containing aggression, which focus on suppressing inflammation.

Summary and conclusions

To summarize, the articles in this collection present partially overlapping as well as strongly diverging opinions on issues dealing with intelligence and adaptive efficiency in a wide range of settings, from social groups to human-machine teams and down to individuals demonstrating performance varying from superior to pathological. The VFEM principle applies at all levels to some degree; from adjusting social policies, correcting individual behavior, and treating pathologies. Understanding is an adaptive strategy within the VFEM scope, expressing integrative operation of two core principles, as follows.

Models represent regularities in the record of sensory inflows and an organism's responses, and vary in scope: from representing contiguous elements in short segments in the record to representing non-contiguous element groupings separated by indefinitely large segments (Yufik, 1998, 2018; Yufik and Sheridan, 2002). Regularities constitute compressible components in the record, with the degree of compression dependent on the types of pressure that drive adaptation. In particular, environmental pressure demands minimization of prediction errors (i.e., VFE) consequent on the organism's decisions, while thermodynamic pressure demands maintaining life-compatible ratios of energy intakes

vs. energy expenditures in the brain producing those decisions. The adaptation-by-learning strategy (recall and compare) subsists on low degrees of compression, limiting adaptation scope to low-complexity contingencies in the organism's vicinity (think of predators chasing preys). By contrast, a uniquely human genetic pressure (i.e., curiosity and the desire to understand) requires unlimited expansions of expressivity over spatial, temporal and complexity dimensions—thus creating an incessant demand for compression and the minimization of complexity (think of formulating theories and aha moments when the simplicity of the solution reveals itself).

Biophysics imposes hard constraints on brain development, limiting the size of the neuronal pool and the ratio of energy supply and expenditure compatible with sustaining life. Complexity and thermodynamic (and metabolic) constraints are intimately linked. For example, the Jarzynski equality tells us immediately, that the more we change our mind—in terms of erasing information—the more energy we consume. Technically, this enables one to associate the complexity of our world models with the metabolic cost of maintaining them in open exchange with the environment. Grasp (abrupt unification of disparate neuronal processes in coherent and self-coordinating structures) aptly responds to all three forms of pressure under complexity and thermodynamic constraints, i.e., grasp mechanisms allow unlimited expansion in the scope of regularities captured in world models, while yielding adequate prediction accuracy at sustainable energy costs. Grasp extracts the essence (the gist) of a situation, enabling predictions at costs that are infinitesimally small in comparison with those the system would be facing without grasp. To fully appreciate the scale of savings, think of 15 moves look-ahead reported by world class masters (Kasparov, 2007). Shannon's (1950) formula puts the number of possible games after 15 half-moves at approximately 2×10^{21}. Making an assumption that a player can evaluate one such possibility per second and can keep this rate up for 30 mins obtains about 2×10^3 evaluations, indicating reduction in the amount of processing on the scale $10^{18} : 1$. Figuratively, grasp confines costly evaluations (reasoning about moves) to the gist of the position held within a hair thin path in a combinatorial ocean that is million times wider than the Pacific. Some articles in this Issue resonate with the above ideas, while some others offer interesting alternatives.

In conclusion, we offer some observations and suggestions for future research in biological and machine intelligence. The history of the latter can be divided into four periods: pebbles, abacus, calculators and computers. Gadgets of the former three types hold only data, while algorithms for manipulating data remain in the mind of the user. The computer revolution was propelled by the realization (John von Neumann) that algorithms can be held alongside data in the same medium. This revolution allowed the delegation of learning to machines, with the temptation to reduce all of higher cognition to algorithmic data manipulation (machine learning). As a result, progress in machine intelligence has relied primarily on advances in the efficiency of data manipulation, which is, in a way orthogonal to that exploited by evolution (human neurons are not faster, smaller or more energy efficient than other species, though there are more of them). The tremendous value produced by machine learning does not change the fact that, in principle, learning machines operate in a context invariant fashion—in familiar conditions—and can only deceive users into ascribing understanding to them while, in fact, having none.

Evolution has explored the adaptation-by-learning route in millions of species and during billions of years since the emergence of life on earth, and ran into a dead end in higher animals. Understanding is a product of a recent evolutionary discovery [which, conceivably, coopted some existing mechanisms (Yufik, 2018, 2021a) that, in about 100,000 years, advanced human civilization from foraging and hunting to launching missiles and sending telescopes to the outer space]. The core mental act of 'merging pieces together' is non-verbalizable but could have given birth to language (Berwick and Chomsky, 2017). The adaptive value of a non-algorithmic "grasp" derives precisely from its ability to overcome inertia and dissolve templates acquired in the course of learning. It is not unreasonable to assume that imparting a modicum of understanding capacity to machines could bring about benefits on a par with or greater than those delivered by the computer revolution.

Technically speaking, the transition from machine learning to machine understanding shifts the research emphasis from representing recognition *via* vector mapping (as in neural nets) to representing relations *via* coordinated vector movement (think of the domino row and associate direction vector with each piece—considering that rotating one vector in the first piece brings about similar rotations in others). Challenges posed by deviating from the von Neumann–Turing architecture and/or designing computable approximations of the ways understanding operates might be stupendous but not insurmountable (Siegelmann, 1999; Yufik, 2002; Traversa and Di Ventra, 2017; Di Ventra and Traversa, 2018; Hylton, 2022). VFEM does not stipulate methods for implementing machine intelligence but constrains the conceptual or computational space for formulating them and establishes a tractable performance metric. Arguably, the problem of machine consciousness is subordinate to that of machine understanding: if understanding is a lens, consciousness acts as an eyelid: one can see when the lid is up and not when it is down (with degrees of clarity depending on the degree of squinting).

A recent book on expert decision making was entitled "Sources of Power" (Klein, 2017), whose title coheres with one of the key insights in a philosophical classic:

> "Quite generally, the familiar, just because it is familiar, is not cognitively understood. The commonest way in which we deceive either ourselves or others about understanding is by assuming something is familiar and accepting it on that account; with all its pros and cons, such knowing never gets anywhere, and it knows not why.
>
> … The analysis of an *idea,* as it used to be carried out was, in fact, nothing else than ridding it of the form in which it had become familiar. …The activity of dissolution is the power and work of the *Understanding,* the most astonishing and mightiest of powers, or rather the absolute power" [Hegel, [1977 (1807)], p. 18].

Harnessing this power can be decisive in securing competitive edge in commerce and defense.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Author's disclaimer

The views expressed in this article are solely those of the authors and do not necessarily represent those of the United States Air Force.

Conflict of interest

Author YY was employed by Virtual Structures Research Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aristotle (2004). *The Metaphysics*. London: Penguin Books.

Barsalou, L. W. (2008). Cognitive and neural contributions to understanding the conceptual system. *Curr. Direct. Psychol. Sci.* 17, 91–95. doi: 10.1111/j.1467-8721.2008.00555.x

Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus* 40/98, 67–88. doi: 10.1515/cpt-2014-0002

Baumberger, C., Beisbart, C., and Brun, G. (2016). "What is understanding? An overview of recent debates in epistemology and philosophy of science," in *Explaining Understanding. New Perspectives from Epistemology and Philosophy of Science*, eds S. Grimm and G. Baumberger (New York, NY: Routledge).

Berkeley, G. [1998 (1734)]. *A Treatise Concerning the Principle of Human Knowledge*. Oxford: Oxford University Press.

Berwick, R. C., and Chomsky, N. (2017). *Why Only US: Language and Evolution.* Cambridge, MA: The MIT Press.

Chase, W. G., Herbert, A., and Simon, H. A. (1973). Perception in chess. *Cog. Psychol.* 4, 55–81. doi: 10.1016/0010-0285(73)90004-2

De Groot, A. (1965). *Though and Choice in Chess*. Hague, Netherlands: Mouton.

de Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

Descartes, R. [1998 (1637)]. *Discourse on Method*. Indianapolis: Hackett Publishing Company, Inc.

Di Ventra, M., and Traversa, F. L (2018). Memcomputing: Leveregaing memory and physics to compute efficiently. *J. Appl. Phys.* 123, 1–18. doi: 10.1063/1.5026506

Freeman, W. J. (2000). *How Brains Make UP Their Minds.* New York, NY: Columbia University Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain (archive). *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Gelepithis, P. A. M. (1986). Conceptions of human understanding: a critical review. *Cogn. Syst.* 1, 295–305.

Gobet, F., and Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychol. Sci.* 7, 52–55. doi: 10.1111/j.1467-9280.1996.tb00666.x

Greco, J. (2014). "Episteme: knowledge and understanding," in *Virtues and Their Vices*, eds K. Timpe and C. A. Boyd (Oxford: Oxford University Press), 285–303.

Grimm, S. R. (2006). Is understanding a species of knowledge? *Br. J. Philos. Sci.* 57, 515–535. doi: 10.1093/bjps/axl015

Grimm, S. R. (2012). The value of understanding. *Philos. Compass* 7, 103–117. doi: 10.1111/j.1747-9991.2011.00460.x

Grimm, S. R. (2014). "Understanding as knowledge of causes," in *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science,* eds A. Fairweather (Berlin: Springer), 329–345.

Hannon, M. (2021). Recent work in the epistemology of understanding. *Am. Philos. Q.* 58, 269–290. doi: 10.2307/48616060

Hegel, W. F. [1977 (1807)]. *Phenomenology of Spirit.* Oxford, NY: Oxford University Press.

Hume, D. [2018 (1739)]. *A Treatise of Human Nature*. London: Penguin Books.

Hylton, T. (2022) Thermodynamic state-machine network. *Entropy*. 24, 744. doi: 10.3390/e24060744

Kant, I. [1990 (1781)] *Critique of Pure Reason*. Buffalo, NY: Prometheus Books.

Kasparov, G. (2007). *How Life Imitates Chess.* Bloomsbury, NY.

Kauffman, S. (2010). *Is the Human Mind Algorithmic?* Available online at: https://www.npr.org/sections/13.7/2010/03/is_the_human_mind_algorithmic_1.html

Khalifa, K. (2013). The role of explanation in understanding. *Br. J. Philos. Sci.* 64, 161–187. doi: 10.1093/bjps/axr057

Klein, G. A. (2017). *Sources of Power: How People Make Decisions*. Boston, The MIT Press.

Kvanvig, J. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge, NY: Cambridge University Press.

Kvanvig, J. (2009). "The value of understanding," in *Epistemic Value,* eds D. Pritchard, A. Millar, and A. Haddock (Oxford, NY: Oxford University Press), 95–11.

Lear, J. (1988). *Aristotle: The Desire to Understand*. Cambridge: Cambridge University Press.

Locke, J. [1996 (1689)]. *An Essay Concerning Human Understanding*. Indianapolis: Hackett Publishing Company, Ltd.

Mehra, J. (1999). *Einstein, Physics and Reality*. Hackensack, NJ: World Scientific Publishing Co.

Moran, R., Pinotsis, D. A., and Friston, K. (2013). Neural masses and fields in dynamic causal modeling. *Front. Comput. Neurosci.* 7, 57. doi: 10.3389/fncom.2013.00057

Moran, R. J., Symmonds, M., Dolan, R. J., and Friston, K. J. (2014). The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. *PLoS Comput. Biol.* 10, e1003422. doi: 10.1371/journal.pcbi.1003422

Newell, A., and Simon, H. A. (1972). Hoboken, NJ: Human Problem Solving. Prentice Hall.

Parr, T., Pezzulo, G., and Friston, K. J. (2022). *The Free Energy Principle in Mind, Brain and Behavior*. Boston: The MIT Press.

Penrose, R. (1997). On understanding understanding. *Int. Stud. Philos. Sci.* 11, 7–20. doi: 10.1080/02698599708573547

Penrose, R. (2016). *The Emperor's New Mind: Concerning Computers, Mind, and the Laws of Physics*. Oxford: Oxford University Press.

Piaget, J. (1974). *Understanding Causality*. New York, NY: Norton Publishing.

Piaget, J. (1978). *Success and Understanding*. Cambridge, MA: Harvard University Press.

Pritchard, D. (2014). "Knowledge and understanding," in *Virtue Epistemology Naturalized*, eds A. Fairweather (New York, NY: Springer), 315–328. doi: 10.1007/978-3-319-04672-3_18

Reshevsky, S. (2012). *Reshevsky on Chess*. San Rafael, NY: Ishi Press International.

Rovelli, C. (2014). *Reality Is Not What It Seems: The Journey to Quantum Gravity*. Riverhead, NY: Riverhead Books.

Russell, B. [1997 (1921)]. *The Analysis of Mind*. London: Routledge.

Searle, J. R. (1990). Is the brain's mind a computer program? *Sci. Am.* 262, 25–31. doi: 10.1038/scientificamerican0190-26

Shannon, C. E. (1950). XXII. Programming a computer for playing chess, The London, Edinburgh, and Dublin Philosophical Magazine. *J. Sci.* 41, 256–275. doi: 10.1080/14786445008521796

Siegelmann, H. T. (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit*. New York, NY: Springer.

Simon, H. A. (1979). *Models of Thought, vol. 1*. New Haven, CT: Yale University Press.

Traversa, F. L., and Di Ventra, M. (2017). Polynomial-time solution of prime factorization and NP-hard problems with digital memcomputing machines. *Chaos.* 27, 1–22. doi: 10.1063/1.4975761

Yufik, Y., Sengupta, B., and Friston, K. 2017. Self-organization in the nervous system. *Front. Syst. Neurosci.* 11, 69. doi: 10.3389/fnsys.2017.00069

Yufik, Y. M. (1998). "Virtual associative networks: a framework for cognitive modeling," in *Brain and Values*, eds K. Pribram (New Jersey: LEA), 109–177. doi: 10.4324/9780203763834-7

Yufik, Y. M. (2002). "How the mind works," in *Proceedings of IEEE World Congress on Computational Intelligence* (Honolulu, HI: IEEE), 2255–2259.

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M. (2018). "Gnostron: A framework for human-like machine understanding," in *IEEE Symp. Computational Intelligence SSCI 2018* (Bangalore, India), 136–145.

Yufik, Y. M. (2019). The understanding capacity and information dynamics in the human brain. *Entropy* 21, 1–38. doi: 10.3390/e21030308

Yufik, Y. M. (2021a). Laws of nature in action, perception and thinking: Comments on "Laws of nature that define biological action and perception" by M. Latash. *Phys. Life Rev.* 36, 9–11. doi: 10.1016/j.plrev.2020.12.003

Yufik, Y. M. (2021b). "Brain functional architecture and human understanding," in *Connectivity and Functional Specialization in the Brain,* eds T. Heinbockel (London: IntechOpen), 48–64. Available online at: https://www.intechopen.com/chapters/74977

Yufik, Y. M., and Friston, K. (2016). Life and understanding: origins of the understanding capacity in self-organizing nervous systems. *Front. Syst. Neurosci.* 10, 98. doi: 10.3389/fnsys.2016.00098

Yufik, Y. M., and Sheridan, T. (2002). Swiss Army Knife and Ockham's Razor: Modeling operator's comprehension in complex dynamic tasks. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* 32, 185–199. doi: 10.1109/TSMCA.2002.1021107

Yufik, Y. M., and Yufik, T. (2018). "Situational understanding," in *Proceeding Seventh International Conference Advances Computing, Communication and Information* (Rome, Italy), 21–27.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 29 September 2021
doi: 10.3389/fnsys.2021.704069

Bacterial Translocation Associates With Aggression in Schizophrenia Inpatients

Chong Wang[1†], *Teng Zhang*[1†], *Lei He*[1], *Ji-Yong Fu*[1], *Hong-Xin Deng*[1], *Xiao-Ling Xue*[2] *and Bang-Tao Chen*[3*]

[1] Department of Psychiatry, Zhumadian Psychiatric Hospital (The Second People's Hospital of Zhumadian), Zhumadian, China, [2] Department of Hematology, The Third Affiliated Hospital of Chongqing Medical University, Chongqing, China, [3] Department of Dermatology, Chongqing University Three Gorges Hospital, Chongqing, China

Objective: Accumulating evidence indicates that inflammation abnormalities may contribute to aggression behaviors in psychotic patients, however, the possible sources of inflammation remain elusive. We aimed to evaluate the associations among aggression, inflammation, and bacterial translocation (BT) in aggression-affected schizophrenia (ScZ) inpatients with 2 weeks of antipsychotics discontinuation.

Methods: Serum specimens collected from 112 aggression and 112 non-aggression individuals with ScZ and 56 healthy adults were used for quantifications of inflammation- or BT-related biomarkers. Aggression severity was assessed by Modified Overt Aggression Scale (MOAS).

Results: Proinflammation phenotype dominated and leaky gut-induced BT occurred only in cases with ScZ with a history of aggression, and the MOAS score positively related to levels of C-reactive protein, interleukin (IL)-6, IL-1β, and tumor necrosis factor-α. Furthermore, serum levels of BT-derived lipopolysaccharide (LPS), as well as LPS-responded soluble CD14, were not only positively correlated with levels of above proinflammation mediators but also the total MOAS score and subscore for aggression against objects or others.

Conclusion: Our results collectively demonstrate the presence of leaky gut and further correlate BT-derived LPS and soluble CD14 to onset or severity of aggression possibly by driving proinflammation response in inpatients with ScZ, which indicates that BT may be a novel anti-inflammation therapeutic target for aggression prophylaxis.

Keywords: schizophrenia, aggression, bacterial translocation, inflammation, association

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Derek L. Buhl,
Takeda, United States
Laura M. Harrison,
Tulane University, United States

***Correspondence:**
Bang-Tao Chen
medisci@163.com

[†]These authors have contributed equally to this work

Received: *01 May 2021*
Accepted: *19 August 2021*
Published: *29 September 2021*

Citation:
Wang C, Zhang T, He L, Fu J-Y, Deng H-X, Xue X-L and Chen B-T (2021) Bacterial Translocation Associates With Aggression in Schizophrenia Inpatients. Front. Syst. Neurosci. 15:704069. doi: 10.3389/fnsys.2021.704069

INTRODUCTION

Schizophrenia (ScZ) is a chronic and heterogeneous psychiatric syndrome characterized by recurrent episodes of acute psychosis alternating with periods of full or partial remission. Globally, ScZ affects ∼1% of the population and occurs mainly in individuals in the late adolescence or early adulthood (Kahn et al., 2015; Charlson et al., 2018). It covers a broad spectrum of clinical symptoms including positive symptoms (delusions, hallucinations, etc.), negative symptoms (anhedonia, social withdrawal, poverty of thought, etc.), and cognitive dysfunction. Current treatment modalities are available only for symptoms mitigation, thus, significant disability,

insupportable psychosocial burdens, and premature mortality are of great concerns (Tiihonen et al., 2017; Stepnicki et al., 2018).

Compared with the general population, inpatients with ScZ are four to seven times more likely to commit aggression acts involving verbal threat, assault, and homicide, which brings a great challenge for both mental health services and public safety (Cho et al., 2019). Aggression is more inclined to be an independent entity. The manifold pathogenesis of aggression in ScZ is complicated by elevated serum C-reactive protein (CRP) and increased ratio of CRP to interleukin (IL)-10, which arouses increasing concerns about the role of systemic inflammation in the onset or severity of aggression in ScZ (Barzilay et al., 2016; Das et al., 2016; Zhang et al., 2017). Inflammation phenotype involves the integration of various pro-/anti-inflammatory cytokines. Interleukin-6, IL-1β, and tumor necrosis factor (TNF)-α are well-proved proinflammatory cytokines responsible for initiation and exacerbation of inflammation, and the serum levels of them were demonstrated to be significantly upregulated in patients with ScZ in most of the related studies (Lesh et al., 2018; Momtazmanesh et al., 2019). Although other cytokines such as interferon (IFN)-γ, IL-4, IL-17, IL-10, and transforming growth factor (TGF)-β were also proved to be linked with ScZ, they may promote or suppress inflammation response in the different subsets of cases with ScZ (Lesh et al., 2018; Momtazmanesh et al., 2019). Crossing blood-brain barrier, the peripheral cytokines precipitate changes in mood and behavior through hypothalamic–pituitary–adrenal axis (Petra et al., 2015; Singh et al., 2019; Misiak et al., 2020), which lays a structural foundation for studying the involvement of inflammation in aggression. However, the mentioned cytokines except serum CRP are rarely profiled and potential sources of peripheral inflammation, with exception of being overweight or lack of dental care, are seldom explored (Fond et al., 2021) in individuals with aggression (Ag)-affected ScZ (ScZ-Ag).

Interestingly, inflammation abnormalities could be caused by alterations in the gut microbiome and the recent evidence from human metabolomics suggested a correlation between enteric dysbacteriosis and dysfunction of neurochemical pathways including inflammation activation underlying aggression in patients with ScZ (Severance et al., 2016; Manchia and Fanos, 2017; Chen et al., 2021; Zeng et al., 2021). Changes in gut microbiota may compromise the integrity of the intestinal tract (leaky gut) and subsequently cause a higher translocation rate of bacterial immunogenic components such as bacterial DNA (BactDNA) and lipopolysaccharide (LPS) from gut into peripheral circulation, which in turn activate immuno-inflammatory signaling (Francés et al., 2007; Martin-Subero et al., 2016). The so-called bacterial translocation (BT) was extensively proved to be correlated with various inflammation-involved diseases and with negative symptoms or neurocognitive impairments in deficit cases with ScZ (Caso et al., 2016; Maes et al., 2019a; Severance et al., 2020). However, the occurrence of leaky gut-related BT and its association with systemic inflammation in ScZ-Ag are poorly investigated.

Taken together, we hypothesize that proinflammation cytokines characterize the aggression behaviors in patients with ScZ and increased intestinal permeability-caused BT is one of the main culprits for the tuning process of inflammation. With regard to this, we determined serum levels of aforementioned inflammation cytokines, leaky gut and BT-related biomarkers, and further assessed the correlations between BT biomarkers and inflammation cytokines or the severity of aggression, in the hope of providing more convincing evidence for BT-derived inflammatory pathogenesis of aggression in ScZ.

MATERIALS AND METHODS

Study Population

The prospective and controlled investigation was conducted in inpatients with ScZ with or without aggression behaviors within 1 week prior to admission during November 2019 to November 2020 in the Second People's Hospital of Zhumadian, a tertiary psychiatric hospital in Henan Province, China. At sample collection, all included inpatients with ScZ were at least 2 weeks of antipsychotics discontinuation. Inpatients with ScZ with the presence of aggression behaviors within 1 week prior to admission and absence of any aggression behaviors during disease course before enrollment were classified into ScZ-Ag and NScZ-Ag groups, respectively.

For comparison, age-, gender-, and body mass index (BMI)-matched healthy volunteers recruited during the same period with no history of psychiatric or medical illness were set as control (Ctrl group) and the ratio of healthy volunteers: cases with ScZ-Ag is 1:2. All the subjects were aged ≥18 years. The diagnosis was made by two board-certified psychiatrists according to the 10th edition of the international classification of diseases (ICD-10) criteria for ScZ. Exclusion criteria included: (a) aggression behaviors not within 1 week prior to admission; (b) pregnant or lactating women; (s) presence of any other psychoses including affective disorder or substance abuse; (d) comorbidity with severe somatic diseases or neurological diseases; (e) comorbidity with other medical conditions such as parenchymal organ-specific diseases, immune-related diseases, hematological diseases, gastrointestinal diseases, and any history of gastrointestinal surgeries; (f) use of systemic corticosteroids, any other immunosuppressive therapy, and oral probiotics in the recent 3 months; (g) inpatients with fever (>37.9°C) or those who were treated with antibiotics, antipyretics or anti-inflammatory medications in the recent 2 weeks. The study was approved by and carried out under the guidelines of the Ethics Committee of the Hospital, and written informed consent was obtained from all the healthy volunteers, the inpatients or the guardians of inpatients (if the patients were unable to sign consent because of poor intelligence) at the time of recruitment.

Subjects Profiles

A structured questionnaire was used to collect data on general sociodemographic variables (age, gender, occupation, education background, ethnicity, height and weight, family income, living circumstance and marriage status), health status (medical history, current medications and family history), and living habits (alcohol intake and smoking pattern) in all the participants. In inpatients with ScZ, the information on specific conditions

including the onset of illness and the type of aggression was inquired.

Clinical Assessments

Modified Overt Aggression Scale (MOAS) was used to characterize aggression behaviors observed within the past 1 week. It involves four subscales and a score from zero to four is assigned for each type of aggression with zero indicating no aggression and higher scores pointing to increasing severity. The score of each subscale is then multiplied by a predefined loading (one for verbal aggression, two for aggression against objects, three for self-aggression, and four for aggression against other people) and the sum of each subscale-weighted score (range 0–40) is referred to the total score. Inpatient with a total score of zero or only having a score of one or more for verbal aggression was classified as being the non-aggressive (Huang et al., 2009). The presence and severity of each psychiatric symptom in cases with ScZ were evaluated by the positive and negative syndrome scale (PANSS) involving positive symptom subscale (seven items), negative symptom subscale (seven items), general psychopathological subscale (16 items), and supplemental items (three items). Each item on the subscale score from one to seven base on the frequency and severity of the symptom (Kelley et al., 2013).

Blood Sampling and Laboratory Detection

Fasting peripheral blood samples were collected from all the subjects at 8:00 a.m. Blood cell count and liver function were examined routinely. The protein levels of indicators assessed by enzyme-linked immunosorbent assay (ELISA) in this study involved CRP (#E007462, 3ABio, Shanghai, China), IL-6 (#E000482, 3ABio, Shanghai, China), IL-1β (#E001772, 3ABio, Shanghai, China), IL-4 (#DG10308H, Dogesce, Beijing, China), IL-10 (#DG10495H, Dogesce, Beijing, China), IL-17 (#DG10431H, Dogesce, Beijing, China), IFN-γ (#C608-01, GenStar, Beijing, China), TNF-α (#489204, Cayman, Michigan, USA), TGF-β (#DG10113H, Dogesce, Beijing, China); leaky gut-related biomarkers[intestinal fatty acid-binding protein (I-FABP, #DFBP20, R&D Systems, Minnesota, USA) and Claudin-3 (#abx250611, Abbexa, Cambridge, UK)]; BT-related biomarkers[LPS (#DG11072H, Dogesce, Beijing, China), soluble CD14 (sCD14, #DC140, R&D Systems, Minnesota, USA), and endotoxin core antibody (EndoCAb, #E013362, 3ABio, Shanghai, China)]. Assays were performed according to the specifications of the manufacturer and the detection limits were in line with the instructions of the manufacturer. Each serum sample was measured in duplicate. All the plates were read by the I Mark™ Micro plate Reader (Bio-Rad, Hercules, California, United States).

Quantification of BactDNA Fragments

Quantification of circulating BactDNA fragments and quality control were performed as described previously (Such et al., 2002; Ericsen et al., 2016). To avoid potentially bacterial contamination of molecular biology reagents, all the specimens were processed in airflow chambers by the same investigator and all the tubes were never exposed to free air. To remove potentially confounding 16S rDNA contamination, six tubes of prepared diethyl pyrocarbonate (DEPC) water were set as negative controls and the processes of water from DNA extraction to quantitative PCR (qPCR) were completely synchronized with those of blood.

Genomic DNA was extracted from 200 μl of serum or DEPC water using QIAmp DNA Blood Minikit (Qiagen, Hilden, Germany) according to the instructions of the manufacturer and DNA was eluted in a 100 μl final volume. BactDNA levels were determined by qPCR in an amplification reaction of 20 μl with forward primer (5′-AGAGGGTGATCGGCCACA-3′) and reverse primer (5′-TGCTGCCTCCCGTAGGAGT-3′), the universal eubacterial primers of a conserved region of 16S rDNA gene (Francés et al., 2004). The amplification conditions for the 59 base pairs of DNA fragments were 95°C for 10 min, followed by 45 cycles at 95°C for 15 s and 60°C for 60 s. Each sample was amplified in triplicate and the BactDNA content was calculated according to a standard curve that generated from serial dilutions of plasmid DNA containing known copy numbers of the template. The final circulating BactDNA concentration was calculated by subtracting the proportion of 16S rDNA copies/μl detected in water controls from those in blood.

Statistical Analyses

Statistical analysis of the data compiled in Excel databank was conducted using SPSS/PC software (Version 19.0 for Windows; SPSS Inc., China). Categorical and continuous variables were expressed as number (%) or mean (M) ± SD, respectively. Normal distribution of raw data was inspected by Kolmogorov–Smirnov tests, and IL-17, IGF-β, and EndoCAb were logarithmically transformed to achieve Gaussian distributions. There were no outliers in MOAS score, PANSS score, cytokines, and bacterial measures by inspection of related boxplots. For comparison of demographic information and clinical characteristics at baseline among groups, Fisher's exact Chi-square test or one-way ANOVA were conducted except specification. Analysis of covariance (ANCOVA) controlling for age, gender, BMI, and course with ScZ was used to analyze cytokines and bacterial measures among the three groups, and Bonferroni's multiple comparison test that can calculate the corrected statistical significance for multiple comparisons was performed for *post-hoc* analysis of pairwise comparisons. Partial correlation analysis controlling for episodes with ScZ, course with ScZ, income levels, marriage status, education background, and occupation was used to determine the relationship between clinical symptoms and inflammation cytokines or bacterial measures. All the tests were two-sided. A $P < 0.05$ was accepted as the cutoff for statistical significance.

RESULTS

Inpatients Characteristics

During the time of study, a total of 528 adult inpatients with ScZ demonstrated a history of aggression behaviors prior to hospitalization. By excluding cases with <2 weeks of antipsychotics discontinuation (56 cases), aggression occurred prior to 1 week time period preceding hospital admission (135 cases), aggression occurred prior to and within 1 week (184

TABLE 1 | Clinic characteristics of all inpatients at baseline.

Items	ScZ-Ag group	NScZ-Ag group	Ctrl group	*P*-value
Case No.	112	112	56	-
Female, n (%)	67 (59.8)	64 (57.1)	35 (62.5)	0.792
Age, mean (SD), years	33.5 (8.4)	34.2 (9.1)	33.8 (8.7)	0.835
BMI, mean (SD), kg/m^2	21.9 (2.4)	21.7 (2.1)	22.3 (1.9)	0.243
Ethnic Han, n (%)	105 (93.8)	109 (97.3)	52 (92.9)	0.336
Low income, n (%)	72 (64.3)	41 (36.7)	17 (30.4)	**0.000**
Living with families, n (%)	91 (81.3)	96 (85.7)	45 (80.4)	0.579
Marriage status, n (%)				
Married	18 (16.1)	33 (29.5)	35 (62.5)	**0.000**
Single	66 (58.9)	35 (31.2)	15 (26.8)	
Divorced	26 (23.1)	37 (33.0)	6 (10.7)	
Widowed	2 (1.9)	7 (6.3)	0 (0)	
Education background, n (%)				
Elementary school and below	82 (73.2)	49 (43.8)	12 (21.4)	**0.000**
Middle and high school	15 (13.4)	39 (34.8)	16 (28.6)	
College and above	15 (13.4)	24 (21.5)	28 (50.0)	
Occupation, n (%)				
Physical labor	21 (18.8)	22 (19.7)	18 (32.1)	**0.000**
Mental labor	13 (11.6)	10 (8.9)	29 (51.8)	
Unemployment	78 (69.6)	80 (71.4)	9 (16.1)	
No. of ScZ episodes[&]	5.6 (2.7)	4.9 (3.2)	NA	0.319
ScZ course, mean (SD), years[&]	7.4 (4.3)	6.9 (4.1)	NA	0.431
Total MOAS score[&]	16.4 (8.2)	1.6 (0.9)	0 (0.0)	**0.000**
Total PANSS score[&]	65.2 (8.3)	63.4 (7.5)	0 (0.0)	0.090

ScZ-Ag, schizophrenia with aggression; NScZ-Ag, schizophrenia without any aggression; BMI, body mass index; MOAS, Modified Overt Aggression Scale; PANSS, positive and negative syndrome scale; NA, not applicable; &, analysis using Student's t-test between ScZ-Ag and NScZ-Ag groups. The meaning of the bold values indicate P< 0.05.

cases), and aggression occurred only within 1 week but met the aforementioned exclusion criteria (41 cases), only 21.2% (112/528) of them [average total MOAS score, mean(SD), 16.4(8.2)] were included in ScZ-Ag group as defined previously. In this study, 112 age-, gender-, and BMI-matched NScZ-Ag inpatients [average total MOAS score, mean(SD), 1.6(0.9)] and 56 healthy volunteers were included. As **Table 1** showed, there was statistical significance in terms of income, marriage, education level, and occupation among the three groups ($P < 0.001$ for all variables). Compared with NScZ-Ag group, more aggression inpatients were single (58.9 vs. 31.2%, $P = 0.010$) and a much higher proportion of aggression cases had low income (64.3 vs. 36.7%, $P = 0.017$) and poor education background (73.2 vs. 43.8%, $P = 0.021$). Between inpatients with ScZ with and without aggression, there was no statistical difference regarding ethnicity, living conditions, occupation distribution, episodes with ScZ, course with ScZ, and total PANSS score ($P > 0.05$ for all the variables).

Inflammation and Severity of Aggression

As shown in **Figure 1**, the results of ANCOVA analysis displayed that there were statistically significant differences between ScZ-Ag, NScZ-Ag, and Ctrl groups in terms of CRP ($F = 75.2$, $P < 0.001$), IL-6 ($F = 102.00$, $P < 0.001$), IL-1β ($F = 37.90$, $P < 0.001$), TNF-α ($F = 450.00$, $P < 0.001$), IL-17 ($F = 7.00$, $P = 0.007$), and TGF-β ($F = 7.55$, $P = 0.008$). Further, *post-hoc* analysis using Bonferroni's multiple comparison test found that none of inflammatory markers differed significantly between NScZ-Ag and Ctrl groups (all $P > 0.05$), while serum levels of CRP, IL-6, IL-1β, and TNF-α dramatically increased approximately two to five times on average in ScZ-Ag group in comparison with NScZ-Ag group (all $P < 0.001$). On partial correlation analysis controlling potential confounders, serum levels of CRP ($r = 0.309$, $P < 0.001$), IL-6 ($r = 0.526$, $P < 0.001$), IL-1β ($r = 0.552$, $P < 0.001$), and TNF-α ($r = 0.517$, $P < 0.001$) were all positively associated with total MOAS score in ScZ-Ag group (**Figure 2**). Altogether, these results indicate that systemic proinflammation response mainly occurs in inpatients with ScZ with aggression behaviors.

BT Determination and Its Association With Inflammation

To explore the source of proinflammation phenotype, BT-related serum biomarkers in all the subjects were measured (**Figure 3**). Regarding biomarkers of "leaky gut" (Claudin-3 and I-FABP), bacterial components (LPS and BactDNA), and LPS-response products (sCD14 and EndoCAb), statistically significant differences between the three groups were observed (all $P < 0.01$) from ANCOVA analysis results. *Post-hoc* analysis showed that only BactDNA titers (11.79 ± 6.97 vs. 7.19 ± 4.76 copies/μl, $P < 0.001$) and sCD14 levels (1.57 ± 1.15 vs. $1.07 \pm 0.61 \times 10^6$ pg/ml, $P < 0.05$) were moderately increased in NScZ-Ag group than Ctrl group, while serum concentrations of Claudin-3 (58.47 ± 13.52 vs. 39.27 ± 9.61 ng/ml, $P < 0.001$), I-FABP (80.47 ± 21.47 vs. 29.56 ± 7.46 pg/ml, $P < 0.001$), LPS (73.51 ± 32.29 vs. 23.16 ± 7.83 pg/ml, $P < 0.001$), sCD14 (3.45 ± 1.39 vs. $1.57 \pm 1.15 \times 10^6$ pg/ml, $P < 0.001$) were significantly higher, EndoCAb concentration (2.18 ± 0.13 vs. 2.23 ± 0.11 $\log_{10}$ MMU/ml, $P < 0.01$) was remarkably lower in ScZ-Ag group than NScZ-Ag group. In ScZ-Ag group (**Table 2**), circulating concentration of LPS was further found to be positively correlated with CRP ($P < 0.001$), IL-1β ($P = 0.001$) and TNF-α ($P = 0.006$), sCD14 was positively associated with CRP ($P < 0.001$), IL-6 ($P = 0.007$), and TNF-α ($P = 0.040$) after controlling potential confounders. These data not only indicate the presence of "leaky gut," but also imply the link that circulating LPS from BT, as well as LPS responded sCD14, might be the important cause synergistically leading to the higher levels of proinflammation mediators observed in inpatients with ScZ with any type of aggression behaviors.

Correlation of BT With Symptoms Dimensions

Partial correlation analyses in inpatients with aggression (**Table 3**) showed that total MOAS score was positively associated with protein levels of circulating LPS ($r = 0.412$, $P = 0.005$) or sCD14 ($r = 0.267$, $P = 0.035$). Regarding the subscale of MOAS, only aggression against objects ($r = 0.406$, $P = 0.006$) or toward others ($r = 0.326$, $P = 0.011$) were found to be correlated positively with circulating LPS, and such associations with the circulating sCD14 were also detected. In addition, results showed

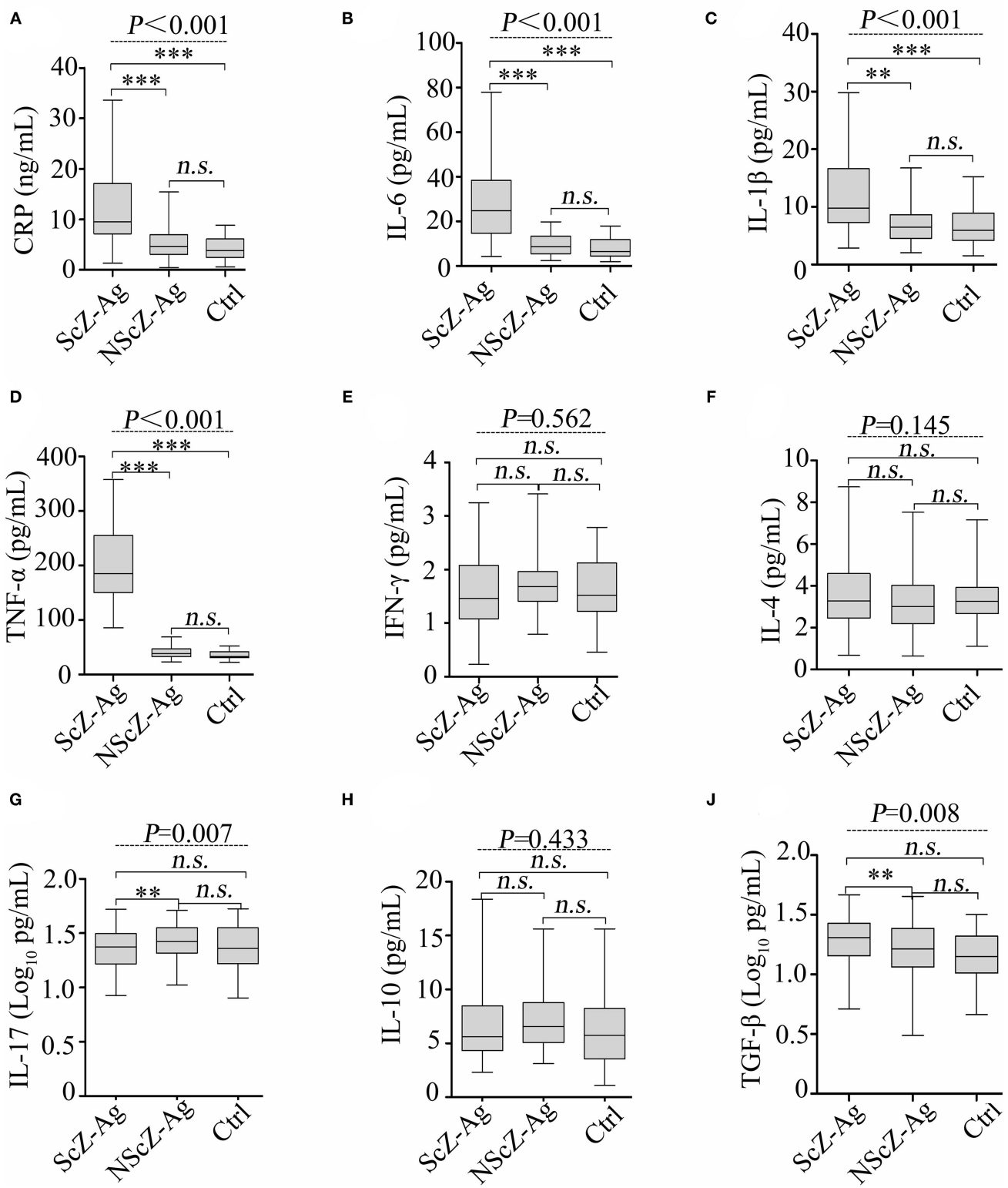


FIGURE 1 | Protein levels of serum CRP **(A)**, IL-6 **(B)**, IL-1β **(C)**, TNF-α **(D)**, IFN-γ **(E)**, IL-4 **(F)**, IL-17 **(G)**, IL-10 **(H)** and TGF-β **(J)** in peripheral blood of subjects. CRP, C-reactive protein; IL, interleukin; TNF, tumor necrosis factor; TGF, transforming growth factor. Data were presented as boxplots. In *post-hoc* analysis using Bonferroni's multiple comparison test, *n.s.* > 0.05, *P < 0.05, **P < 0.01, ***P < 0.001 analysis using ANCOVA.

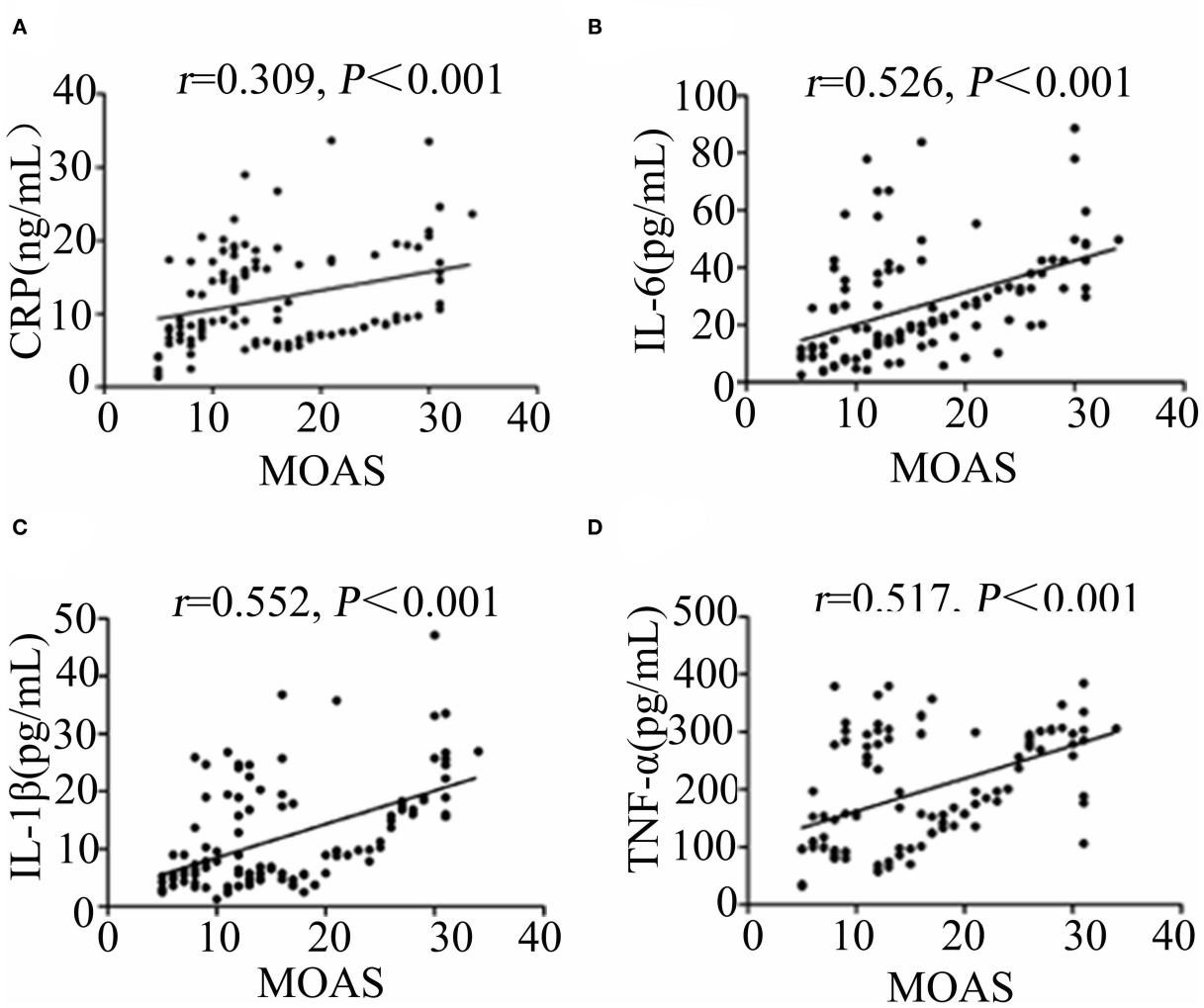


FIGURE 2 | Partial correlation analyses of MOAS with CRP **(A)**, IL-6 **(B)**, IL-1β **(C)**, and TNF-α **(D)** in ScZ-Ag group.

statistically significant association between positive PANSS and circulating LPS ($r = 0.298$, $P = 0.023$) or sCD14 ($r = 0.315$, $P = 0.015$). Altogether, the data further suggest that the increased protein levels of LPS or sCD14 in peripheral blood potentially initiate aggression behaviors in inpatients with ScZ *via* exacerbating the severity of systemic inflammation.

DISCUSSION

Aggression can attack individuals with or without psychosis. It is one of the top 20 causes of disabilities worldwide that is present in 15.3–53.2% of inpatients with ScZ in China (Zhou et al., 2016). Growing evidence demonstrate that the serious public health problem is the resultant of pro-/anti-inflammation imbalance, since some inflammation cytokines were proved to be involved in the pathogenesis of ScZ (Müller et al., 2015; Petrikis et al., 2015; Momtazmanesh et al., 2019; Feng et al., 2020; Park and Miller, 2020). However, the role of and alterations in these cytokines may be variable in different stratifications of ScZ, antipsychotic drugs used or presence of aggression behaviors is a case (Petrikis et al., 2017; Momtazmanesh et al., 2019). This study was the first to focus on aggression-affected inpatients with ScZ with at least 2 weeks of antipsychotics discontinuation. Different from previous studies (Miller et al., 2011; de Witte et al., 2014; Momtazmanesh et al., 2019), our results from Bonferroni's multiple comparison tests demonstrated no difference in inflammation phenotype between inpatients with ScZ without aggression and healthy controls. The contradictory results may be attributed to differences in statistical analysis methods used and the specific enrolled participants without any aggression behaviors during the disease course, which further verifies the inconsistent conclusions regarding inflammatory phenotypes in ScZ (Momtazmanesh et al., 2019). In sharp contrast, dramatical elevations of CRP, IL-6, IL-1β, and TNF-α were not only observed in inpatients with ScZ with aggression, but the elevated cytokines also correlated positively to the severity of aggression measured by MOAS score that is partly similar to the previous reports (Petrikis et al., 2015; Zhang et al.,

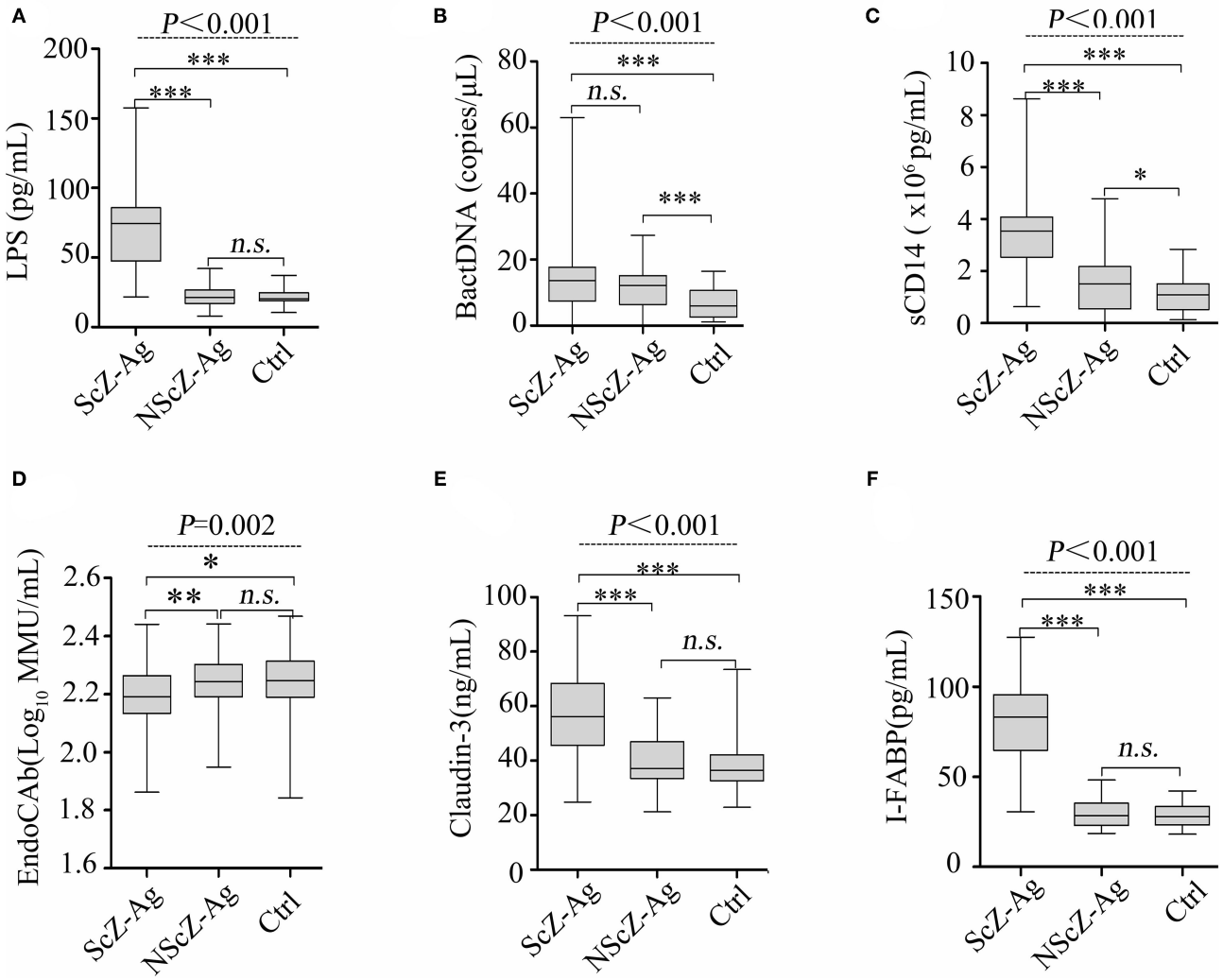


FIGURE 3 | Serum levels of LPS **(A)**, BactDNA **(B)**, sCD14 **(C)**, EndoCAb **(D)**, Claudin-3 **(E)** and I-FABP **(F)** among the three groups. LPS, lipopolysaccharide; BactDNA, bacterial DNA; sCD14, soluble CD14; EndoCAb, endotoxin core antibody; I-FABP, intestinal fatty acid-binding protein. Data were presented as boxplots. In *post-hoc* analysis using Bonferroni's multiple comparison test, *n.s.* > 0.05, *P < 0.05, ***P < 0.001 analysis using ANCOVA.

2017; Orsolini et al., 2018; Momtazmanesh et al., 2019; Fond et al., 2021). Li et al. demonstrated positive correlations between higher plasma IL-17 or TGF-β1 and severity of aggression in patients with ScZ (Li et al., 2016), however, we found slightly lower serum IL-17 and higher serum TGF-β1 in individuals with ScZ with aggression as compared with those without aggression. These findings indicate the need for additional research to confirm the role of IL-17 and TGF-β1 in aggression onset. Among the functional redundancies of IL-6, IL-1β, and TNF-α, these proinflammation mediators potentially drive aggression in a sophisticated and coordinated network. These data collectively suggest the contributory role of systemic proinflammation in the occurrence of aggression in ScZ.

Gastrointestinal source of proinflammation was unveiled in deficit ScZ, and the leaky gut was identified as one of the prerequisites for the inflammatory pathophysiology (Severance et al., 2012, 2016; Barber et al., 2019; Cihákova et al., 2019; Maes et al., 2019b). Transcellular integrity, paracellular adherens, and tight junctions are demonstrated universally to be the structural basis for maintaining normal intestinal permeability. Permeability-related biomarkers such as I-FABP and Claudin-3 present at high levels in peripheral blood can reliably reflect the occurrence of leaky gut as they are released into circulation by enterocytes when intestinal epitheliums are compromised (Barmeyer et al., 2017). The evidence that remarkable increases in serum levels of I-FABP and Claudin-3 only in cases with ScZ with aggression behaviors in this study indicates the possible role of increased intestinal permeability in the onset of proinflammation-driven aggression that has not been previously reported.

Correspondingly, the peripheral blood concentration of LPS was significantly higher in an aggression-affected group with

TABLE 2 | Relations of bacterial translocation markers to cytokines in ScZ-Ag group.

BT markers \ Cytokines	CRP		IL-6		IL-1β		TNF-α	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*
LPS	0.713	**0.000***	0.212	0.298	0.627	**0.001***	0.583	**0.006***
BactDNA	0.201	0.329	0.196	0.472	0.098	0.592	0.302	0.129
sCD14	0.826	**0.000***	0.509	**0.007***	0.056	0.613	0.341	**0.040***
EndoCAb	−0.239	0.269	−0.117	0.523	−0.049	0.617	−0.316	0.084
I-FABP	0.112	0.564	0.082	0.613	0.067	0.627	0.298	0.158
Claudin-3	0.257	0.218	0.286	0.182	0.318	0.065	0.125	0.499

**P < 0.05. Analyses using partial correlation analysis. The meaning of the bold values indicate P< 0.05.*

TABLE 3 | Correlations of Lipopolysaccharide or sCD14 with severity aggression.

Items	LPS		sCD14	
	r	*P*	*r*	*P*
Total MOAS score	0.412	**0.005***	0.267	**0.035***
Verbal aggression	0.198	0.263	0.154	0.299
Aggression against objects	0.406	**0.006***	0.397	**0.008***
Self-aggression	−0.054	0.512	−0.067	0.476
Aggression toward others	0.326	**0.011***	0.256	**0.042***
Total PANSS score	0.068	0.477	0.118	0.364
Positive	0.298	**0.023***	0.315	**0.015***
Negative	−0.136	0.317	−0.098	0.418
General	0.049	0.574	0.009	0.832

**P < 0.05. Analyses using partial correlation analysis. The meaning of the bold values indicate P< 0.05.*

ScZ as compared with the non-aggression. Translocating LPS links with an exacerbation of inflammation response (Panpetch et al., 2020) and the following correlation analysis also showed positive correlativity between the circulating concentrations of proinflammation mediators and LPS, and also LPS responded sCD14. Furthermore, serum levels of both the LPS and sCD14 were found to be related to specific aggression behaviors (aggression against objects or toward others) or psychotic symptoms (positive PANSS). As LPS-specific host response, sCD14 circulates at high levels in the serum and interacts with translocating LPS to stimulate antigen-presenting cells *via* toll-like receptor 4 (TLR4) signaling (Tsukamoto et al., 2018). Under bacteria or LPS challenge, vascular endothelial cells and perivascular mast cells have been reported to express abundant TLR4, thus, initiating the production of inflammation cytokines (Zeuke et al., 2002). On the other hand, decreased host EndoCAb in peripheral blood failed to bind and clear LPS from circulation, which ensures a high serum level of LPS for a long time and subsequently maintains systemic inflammation (Kyosiimire-Lugemwa et al., 2020). It is also worth noting that serum BactDNA loads in cases with ScZ with aggression may have little effect on inflammation state given the results from correlation analysis and differential expressions of BactDNA among cases with or without aggression. We can only speculate that serum BactDNA loads quantified by qPCR likely underestimate the presence of BactDNA within whole blood and corresponding perturbation of inflammation markers may be transient. Collectively, these findings emphasize the implication of translocating LPS as well as sCD14 in the systemic inflammation response, and thus, argue for the potential causative relationship between BT and onset of aggression in ScZ.

Largely due to the failures of interpersonal inference, to develop a proper theory of mind or in sensory attenuation, ScZ was identified as one of the emotion recognition disorders (Demekas et al., 2020). Emotion recognition has also been suggested to underlie aggression in individuals with ScZ (Acland et al., 2021); however, that may be decreased by elevated low-grade inflammation (Balter et al., 2018, 2021). In this study, evidence that systemic proinflammation potentially initiated by serum LPS correlated with aggression severity in inpatients with ScZ implies the possible contribution role of serum LPS to aggressive behaviors *via* emotion misrecognition that has important implications for integrated treatments of aggression.

Unfortunately, at least five limitations exist in our study. At first, a structured clinical interview to determine the clinical diagnosis of subjects was not performed. Second, only single samples from participants in a single center were obtained, within-subject verification of related biomarkers and replication procedures in larger study populations from multicenter are expected. Third, higher circulating BactDNA load was observed in NScZ-Ag group compared with the healthy group (**Figure 3B**), while correlation analysis between BactDNA and inflammation in NScZ-Ag group was not conducted as inflammation cytokines did not differ between the two groups (**Figure 1**), thereby perplexing the function of BactDNA in pathogenesis with ScZ. Furthermore, as with all case-controlled clinical studies, present data failed to adequately explain the causal relationship between BT-caused inflammation response and aggression in ScZ, related animal experiments are expected for ethical considerations. At last, the molecular mechanism by which translocating LPS promotes systemic inflammation and thus drives aggression remains to be further investigated.

CONCLUSION

In conclusion, this study verifies mainly the presence of leaky gut-caused BT and further correlates BT-derived LPS and

soluble CD14 to the severity of aggression possibly by driving proinflammation response in cases with ScZ with aggression. These observations collectively indicate that BT may be a novel anti-inflammation therapeutic target for aggression prophylaxis and improving disease outcomes in patients with ScZ with aggression against objects and others.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Second People's Hospital of Zhumadian. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CW and TZ: contributed equally to the manuscript. J-YF and B-TC: methodology, supervision, visualization, and writing—review and editing. CW and H-XD: clinical assessment, data curation, investigation, and writing—original draft. LH and X-LX: formal analysis and statistical analysis. All authors have contributed to and have approved the final manuscript.

FUNDING

This work was funded by the grant from the National Natural Science Foundation of China (82003337) and China Postdoctoral Science Foundation (2020M683268).

ACKNOWLEDGMENTS

We thank Prof Bing Cao (School of Public Health, Peking University) for her help in statistics and language editing.

REFERENCES

Acland, E. L., Jambon, M., and Malti, T. (2021). Children's emotion recognition and aggression: a multi-cohort longitudinal study. *Aggress. Behav.* doi: 10.1002/ab.21989. [Epub ahead of print].

Balter, L., Hulsken, S., Aldred, S., Drayson, M. T., Higgs, S., Veldhuijzen van Zanten, J., et al. (2018). Low-grade inflammation decreases emotion recognition–evidence from the vaccination model of inflammation. *Brain Behav. Immun.* 73, 216–221. doi: 10.1016/j.bbi.2018.05.006

Balter, L. J., Raymond, J. E., Aldred, S., Higgs, S., and Bosch, J. A. (2021). Age, BMI, and inflammation: associations with emotion recognition. *Physiol. Behav.* 232:113324. doi: 10.1016/j.physbeh.2021.113324

Barber, G. S., Sturgeon, C., Fasano, A., Cascella, N. G., Eaton, W. W., et al. (2019). Elevated zonulin, a measure of tight-junction permeability, may be implicated in schizophrenia. *Schizophr. Res.* 211, 111–112. doi: 10.1016/j.schres.2019.07.006

Barmeyer, C., Fromm, M., and Schulzke, J. D. (2017). Active and passive involvement of claudins in the pathophysiology of intestinal inflammatory diseases. *Pflugers Arch.* 469,15–26. doi: 10.1007/s00424-016-1914-6

Barzilay, R., Lobel, T., Krivoy, A., Shlosberg, D., Weizman, A., and Katz, N. (2016). Elevated C-reactive protein levels in schizophrenia inpatients is associated with aggressive behavior. *Eur. Psychiatry* 31, 8–12. doi: 10.1016/j.eurpsy.2015.09.461

Caso, J. R., Balanzá-Martínez, V., Palomo, T., and García-Bueno, B. (2016). The microbiota and gut-brain axis: contributions to the immunopathogenesis of schizophrenia. *Curr. Pharm. Des.* 22, 6122–6133. doi: 10.2174/1381612822666160906160911

Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., et al. (2018). Global epidemiology and burden of schizophrenia: findings from the global burden of disease study 2016. *Schizophr. Bull.* 44, 1195–1203. doi: 10.1093/schbul/sby058

Chen, X., Xu, J., Wang, H., Luo, J., Wang, Z., Chen, G., et al. (2021). Profiling the differences of gut microbial structure between schizophrenia patients with and without violent behaviors based on 16S rRNA gene sequencing. *Int. J. Legal Med.* 135, 131–141. doi: 10.1007/s00414-020-02439-1

Cho, W., Shin, W. S., An, I., Bang, M., Cho, D. Y., and Lee, S. H. (2019). Biological aspects of aggression and violence in schizophrenia. *Clin. Psychopharmacol. Neurosci.* 17, 475–486. doi: 10.9758/cpn.2019.17.4.475

Cihánková, D., Eaton, W. W., Talor, M. V., Harkus, U. H., and Demyanovich, H., Rodriguez, K., et al. (2019). Gut permeability and mimicry of the Glutamate Ionotropic Receptor NMDA type Subunit Associated with protein 1 (GRINA) as potential mechanisms related to a subgroup of people with schizophrenia with elevated antigliadin antibodies (AGA IgG). *Schizophr. Res.* 208, 414–419. doi: 10.1016/j.schres.2019.01.007

Das, S., Deuri, S. K., Sarmah, A., Pathak, K., Baruah, A., Sengupta, S., et al. (2016). Aggression as an independent entity even in psychosis-the role of inflammatory cytokines. *J. Neuroimmunol.* 292, 45–51. doi: 10.1016/j.jneuroim.2016.01.012

de Witte, L., Tomasik, J., Schwarz, E., Guest, P. C., Rahmoune, H., Kahn, R. S., et al. (2014). Cytokine alterations in first-episode schizophrenia patients before and after antipsychotic treatment. *Schizophr. Res.* 154, 23–29. doi: 10.1016/j.schres.2014.02.005

Demekas, D., Parr, T., and Friston, K. J. (2020). An investigation of the free energy principle for emotion recognition. *Front. Comput. Neurosci.* 14:30. doi: 10.3389/fncom.2020.00030

Ericsen, A. J., Lauck, M., Mohns, M. S., DiNapoli, S. R., Mutschler, J. P., Greene, J. M., et al. (2016). Microbial translocation and inflammation occur in hyperacute immunodeficiency virus infection and compromise host control of virus replication. *PLoS Pathog.* 12:e1006048. doi: 10.1371/journal.ppat.1006048

Feng, T., McEvoy, J. P., and Miller, B. J. (2020). Longitudinal study of inflammatory markers and psychopathology in schizophrenia. *Schizophr. Res.* 224, 58–66. doi: 10.1016/j.schres.2020.10.003

Fond, G., Sunhary de, Verville, P. L., Richieri, R., Etchecopar-Etchart, D., Korchia, T., et al. (2021). Redefining peripheral inflammation signature in schizophrenia based on the real-world FACE-SZ cohort. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 111:110335. doi: 10.1016/j.pnpbp.2021.110335

Francés, R., Benlloch, S., Zapater, P., González, J. M., Lozano, B., Muñoz, C., et al. (2004). A sequential study of serum bacterial DNA in patients with advanced cirrhosis and ascites. *Hepatology* 39, 484–491. doi: 10.1002/hep.20055

Francés, R., González-Navajas, J. M., Zapater, P., Muñoz, C., Caño, R., Pascual, S., et al. (2007). Translocation of bacterial DNA from Gram-positive microorganisms is associated with a species-specific inflammatory response in serum and ascitic fluid of patients with cirrhosis. *Clin. Exp. Immunol.* 150, 230–237. doi: 10.1111/j.1365-2249.2007.03494.x

Huang, H. C., Wang, Y. T., Chen, K. C., Yeh, T. L., Lee, I. H., Chen, P. S., et al. (2009). The reliability and validity of the Chinese version of the modified overt aggression scale. *Int. J. Psychiatry Clin. Pract.* 13, 303–306. doi: 10.3109/13651500903056533

Kahn, R. S., Sommer, I. E., Murray, R. M., Meyer-Lindenberg, A., Weinberger, D. R., Cannon, T. D., et al. (2015). Schizophrenia. *Nat. Rev. Dis. Primers* 1:15067. doi: 10.1038/nrdp.2015.67

Kelley, M. E., White, L., Compton, M. T., and Harvey, P. D. (2013). Subscale structure for the Positive and Negative Syndrome Scale (PANSS): a proposed solution focused on clinical validity. *Psychiatry Res.* 205, 137–142. doi: 10.1016/j.psychres.2012.08.019

Kyosiimire-Lugemwa, J., Anywaine, Z., Abaasa, A., Levin, J., Gombe, B., Musinguzi, K., et al. (2020). Effect of stopping cotrimoxazole preventive therapy on microbial translocation and inflammatory markers among human immunodeficiency virus-infected ugandan adults on antiretroviral therapy: the COSTOP trial immunology substudy. *J. Infect. Dis.* 222, 381–390. doi: 10.1093/infdis/jiz494

Lesh, T. A., Careaga, M., Rose, D. R., McAllister, A. K., Van de Water, J., Carter, C. S., et al. (2018). Cytokine alterations in first-episode schizophrenia and bipolar disorder: relationships to brain structure and symptoms. *J. Neuroinflammation* 15:165. doi: 10.1186/s12974-018-1197-2

Li, H., Zhang, Q., Li, N., Wang, F., Xiang, H., Zhang, Z., et al. (2016). Plasma levels of Th17-related cytokines and complement C3 correlated with aggressive behavior in patients with schizophrenia. *Psychiatry Res.* 246, 700–706. doi: 10.1016/j.psychres.2016.10.061

Maes, M., Sirivichayakul, S., Kanchanatawan, B., and Vodjani, A. (2019a). Breakdown of the paracellular tight and adherens junctions in the gut and blood brain barrier and damage to the vascular barrier in patients with deficit schizophrenia. *Neurotox. Res.* 36, 306–322. doi: 10.1007/s12640-019-00054-6

Maes, M., Sirivichayakul, S., Kanchanatawan, B., and Vodjani, A. (2019b). Upregulation of the intestinal paracellular pathway with breakdown of tight and adherens junctions in deficit schizophrenia. *Mol. Neurobiol.* 56, 7056–7073. doi: 10.1007/s12035-019-1578-2

Manchia, M., and Fanos, V. (2017). Targeting aggression in severe mental illness: the predictive role of genetic, epigenetic, and metabolomic markers. *Progr. Neuro Psychopharmacol. Biol. Psychiatry* 77, 32–41. doi: 10.1016/j.pnpbp.2017.03.024

Martin-Subero, M., Anderson, G., Kanchanatawan, B., Berk, M., and Maes, M. (2016). Comorbidity between depression and inflammatory bowel disease explained by immune-inflammatory, oxidative, and nitrosative stress; tryptophan catabolite; and gut-brain pathways. *CNS Spectr.* 21, 184–198. doi: 10.1017/S1092852915000449

Miller, B. J., Buckley, P., Seabolt, W., Mellor, A., and Kirkpatrick, B. (2011). Meta-analysis of cytokine alterations in schizophrenia: clinical status and antipsychotic effects. *Biol. Psychiatry* 70, 663–671. doi: 10.1016/j.biopsych.2011.04.013

Misiak, B., Łoniewski, I., Marlicz, W., Frydecka, D., Szulc, A., Rudzki, L., et al. (2020). The HPA axis dysregulation in severe mental illness: can we shift the blame to gut microbiota? *Progr. Neuro Psychopharmacol. Biol. Psychiatry* 102:109951. doi: 10.1016/j.pnpbp.2020.109951

Momtazmanesh, S., Zare-Shahabadi, A., and Rezaei, N. (2019). Cytokine alterations in schizophrenia: an updated review. *Front. Psychiatry* 10:892. doi: 10.3389/fpsyt.2019.00892

Müller, N., Weidinger, E., Leitner, B., and Schwarz, M. J. (2015). The role of inflammation in schizophrenia. *Front. Neurosci.* 9:372. doi: 10.3389/fnins.2015.00372

Orsolini, L., Sarchione, F., Vellante, F., Fornaro, M., Matarazzo, I., Martinotti, G., et al. (2018). Protein-C reactive as biomarker predictor of schizophrenia phases of illness? a systematic review. *Curr. Neuropharmacol.* 16, 583–606. doi: 10.2174/1570159X16666180119144538

Panpetch, W., Hiengrach, P., Nilgate, S., Tumwasorn, S., Somboonna, N., Wilantho, A., et al. (2020). Additional Candida albicans administration enhances the severity of dextran sulfate solution induced colitis mouse model through leaky gut-enhanced systemic inflammation and gut-dysbiosis but attenuated by Lactobacillus rhamnosus L34. *Gut Microbes* 11, 465–480. doi: 10.1080/19490976.2019.1662712

Park, S., and Miller, B. J. (2020). Meta-analysis of cytokine and C-reactive protein levels in high-risk psychosis. *Schizophr. Res.* 226, 5–12. doi: 10.1016/j.schres.2019.03.012

Petra, A. I., Panagiotidou, S., Hatziagelaki, E., Stewart, J. M., Conti, P., and Theoharides, T. C. (2015). Gut-microbiota-brain axis and its effect on neuropsychiatric disorders with suspected immune dysregulation. *Clin. Ther.* 37, 984–995. doi: 10.1016/j.clinthera.2015.04.002

Petrikis, P., Voulgari, P. V., Tzallas, A. T., Archimandriti, D. T., Skapinakis, P., and Mavreas, V. (2015). Cytokine profile in drug-naïve, first episode patients with psychosis. *J. Psychosom. Res.* 79, 324–327. doi: 10.1016/j.jpsychores.2015.06.011

Petrikis, P., Voulgari, P. V., Tzallas, A. T., Boumba, V. A., Archimandriti, D. T., Zambetas, D., et al. (2017). Changes in the cytokine profile in first-episode, drug-naïve patients with psychosis after short-term antipsychotic treatment. *Psychiatry Res.* 256, 378–383. doi: 10.1016/j.psychres.2017.07.002

Severance, E. G., Alaedini, A., Yang, S., Halling, M., Gressitt, K. L., Stallings, C. R., et al. (2012). Gastrointestinal inflammation and associated immune activation in schizophrenia. *Schizophr. Res.* 138, 48–53. doi: 10.1016/j.schres.2012.02.025

Severance, E. G., Dickerson, F., and Yolken, R. H. (2020). Complex gastrointestinal and endocrine sources of inflammation in schizophrenia. *Front. Psychiatry* 11:549. doi: 10.3389/fpsyt.2020.00549

Severance, E. G., Yolken, R. H., and Eaton, W. W. (2016). Autoimmune diseases, gastrointestinal disorders and the microbiome in schizophrenia: more than a gut feeling. *Schizophr. Res.* 176, 23–35. doi: 10.1016/j.schres.2014.06.027

Singh, L., Kaur, A., Bhatti, M. S., and Bhatti, R. (2019). Possible molecular mediators involved and mechanistic insight into fibromyalgia and associated co-morbidities. *Neurochem. Res.* 44, 1517–1532. doi: 10.1007/s11064-019-02805-5

Stepnicki, P., Kondej, M., and Kaczor, A. A. (2018). Current concepts and treatments of schizophrenia. *Molecules* 23:2087. doi: 10.3390/molecules23082087

Such, J., Francés, R., Muñoz, C., Zapater, P., Casellas, J. A., and Cifuentes, A. (2002). Detection and identification of bacterial DNA in patients with cirrhosis and culture-negative, nonneutrocytic ascites. *Hepatology* 36, 135–141. doi: 10.1053/jhep.2002.33715

Tiihonen, J., Mittendorfer-Rutz, E., Majak, M., Mehtälä, J., Hoti, F., Jedenius, E., et al. (2017). Real-world effectiveness of antipsychotic treatments in a nationwide cohort of 29 823 patients with schizophrenia. *JAMA Psychiatry* 74, 686–693. doi: 10.1001/jamapsychiatry.2017.1322

Tsukamoto, H., Takeuchi, S., Kubota, K., Kobayashi, Y., Kozakai, S., Ukai, I., et al. (2018). Lipopolysaccharide (LPS)-binding protein stimulates CD14-dependent Toll-like receptor 4 internalization and LPS-induced TBK1-IKKε-IRF3 axis activation. *J. Biol. Chem.* 293, 10186–10201. doi: 10.1074/jbc.M117.796631

Zeng, C., Yang, P., Cao, T., Gu, Y., Li, N., Zhang, B., et al. (2021). Gut microbiota: An intermediary between metabolic syndrome and cognitive deficits in schizophrenia. *Progr. Neuro Psychopharmacol. Biol. Psychiatry* 106:110097. doi: 10.1016/j.pnpbp.2020.110097

Zeuke, S., Ulmer, A. J., Kusumoto, S., Katus, H. A., and Heine, H. (2002). TLR4-mediated inflammatory activation of human coronary artery endothelial cells by LPS. *Cardiovasc. Res.* 56, 126–134. doi: 10.1016/S0008-6363(02)00512-6

Zhang, Q., Hong, W., Li, H., Peng, F., Wang, F., Li, N., et al. (2017). Increased ratio of high sensitivity C-reactive protein to interleukin-10 as a potential peripheral biomarker of schizophrenia and aggression. *Int. J. Psychophysiol.* 114, 9–15. doi: 10.1016/j.ijpsycho.2017.02.001

Zhou, J. S., Zhong, B. L., Xiang, Y. T., Chen, Q., Cao, X. L., Correll, C. U., et al. (2016). Prevalence of aggression in hospitalized patients with schizophrenia in China: a meta-analysis. *Asia Pac. Psychiatry* 8, 60–69. doi: 10.1111/appy.12209

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 13 October 2021
doi: 10.3389/fnsys.2021.669810

Trust as Extended Control: Human-Machine Interactions as Active Inference

Felix Schoeller[1,2], Mark Miller[3,4], Roy Salomon[2] and Karl J. Friston[5]*

[1] Massachusetts Institute of Technology, Cambridge, MA, United States, [2] Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan, Israel, [3] Center for Human Nature, Artificial Intelligence and Neuroscience, Hokkaido University, Sapporo, Japan, [4] Department of Informatics, University of Sussex, Brighton, United Kingdom, [5] Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

In order to interact seamlessly with robots, users must infer the causes of a robot's behavior–and be confident about that inference (and its predictions). Hence, trust is a necessary condition for human-robot collaboration (HRC). However, and despite its crucial role, it is still largely unknown how trust emerges, develops, and supports human relationship to technological systems. In the following paper we review the literature on trust, human-robot interaction, HRC, and human interaction at large. Early models of trust suggest that it is a trade-off between benevolence and competence; while studies of human to human interaction emphasize the role of shared behavior and mutual knowledge in the gradual building of trust. We go on to introduce a model of trust as an agent' best explanation for reliable sensory exchange with an extended motor plant or partner. This model is based on the cognitive neuroscience of active inference and suggests that, in the context of HRC, trust can be casted in terms of virtual control over an artificial agent. Interactive feedback is a necessary condition to the extension of the trustor's perception-action cycle. This model has important implications for understanding human-robot interaction and collaboration–as it allows the traditional determinants of human trust, such as the benevolence and competence attributed to the trustee, to be defined in terms of hierarchical active inference, while vulnerability can be described in terms of information exchange and empowerment. Furthermore, this model emphasizes the role of user feedback during HRC and suggests that boredom and surprise may be used in personalized interactions as markers for under and over-reliance on the system. The description of trust as a sense of virtual control offers a crucial step toward grounding human factors in cognitive neuroscience and improving the design of human-centered technology. Furthermore, we examine the role of shared behavior in the genesis of trust, especially in the context of dyadic collaboration, suggesting important consequences for the acceptability and design of human-robot collaborative systems.

Keywords: trust, control, active inference, human-robot interaction, cobotics, extended mind hypothesis, human computer interaction

OPEN ACCESS

Edited by:
Mikhail Lebedev, Duke University, United States

Reviewed by:
Victor de Lafuente, National Autonomous University of Mexico, Mexico
Alessandro Umbrico, Institute of Cognitive Sciences and Technologies, Italian National Research Council, Italy
Roel Stephan Pieters, Tampere University of Technology, Finland
Andrea Orlandini, National Research Council (CNR), Italy

****Correspondence:***
Felix Schoeller
felixsch@mit.edu

Received: *19 February 2021*
Accepted: *16 August 2021*
Published: *13 October 2021*

Citation:
Schoeller F, Miller M, Salomon R and Friston KJ (2021) Trust as Extended Control: Human-Machine Interactions as Active Inference. Front. Syst. Neurosci. 15:669810. doi: 10.3389/fnsys.2021.669810

INTRODUCTION

Technology greatly extends the scope of human control, and allows our species to thrive by engineering (predictable) artificial systems to replace (uncertain) natural events (Pio-Lopez et al., 2016). Navigating and operating within the domain of regularities requires considerably less motor and cognitive effort (e.g., pressing a switch to lift heavy weights) and less perceptual and attentional resources (Brey, 2000); thereby increasing the time and energy available for other activities. However, the inherent complexity of technological systems invariably leads to a state of "epistemic vulnerability," whereby the internal dynamics of the system are hidden to the user and, crucially, must be inferred from the observer via the behavior of the system. Indeed, current misgivings about machine learning rest upon the issue of explainability and interpretability namely, the extent to which a user can understand what is going on "under the hood" (Došilović et al., 2018). By epistemic vulnerability here we mean that the user relies on inference to understand the machine–what the machine does, how it does it, how its actions change given context, etc. Critically, the lack of opacity of these processes may give rise to suspicions and qualms regarding the agent's goals. What factors influence trust during human-robot interaction, and how does human inference modulate the continuous information exchange in human-computer systems? It is widely recognized that trust is a precondition to (successful) human-machine interactions (Lee and See, 2004; Sheridan, 2019a). However, despite great effort from researchers in the field, we still lack a computational understanding of the role of trust in successful human interactions with complex technological systems. Here, we review contemporary theories of trust and their associated empirical data in the context of human-machine interaction. Drawing on the literature in cognitive science of active inference (Friston et al., 2006), control (Sheridan, 2019b), and hierarchical perception-action cycles (Salge and Polani, 2017), we introduce a cross-disciplinary framework of trust–modeled as a sense of *virtual control*. To understand the role of trust in robotics, we first present a brief overview of basic cognitive functions, focusing on the organization of motor control. We then explain the fundamental components of trust–in terms of active inference–and conclude with some remarks about the emergence and development of trust in the context of dyadic human-robot collaboration (HRC), which we take as a good use case for this approach to trust.

SURPRISE MINIMIZING AGENTS

From the standpoint of contemporary cognitive neuroscience, perception and action are means for living organisms to reduce their surprise (i.e., acquire information) about (past, current, and future) states of the world (Friston et al., 2006). The brain according to this framework is considered to be a constructive, statistical organ that continuously generates hypotheses (i.e., beliefs) to predict the most likely causes of the sensory data it encounters (i.e., sensations). These predictions then guide behavior accordingly in a top-down fashion (Gregory, 1980). Various unifying and complementary theories have been proposed to describe this process (e.g., the free energy principle, active inference, predictive processing, dynamic logic, and the Bayesian brain hypothesis). Three fundamental brain functions are defined as follows: (1) perception senses change in the surroundings, (2) cognition predicts the consequences of change, and (3) action controls the causes of change. This tripartition is reflected in the hierarchical functional architecture of brain systems (Kandel et al., 2000), speaking to the brain as an engine of prediction ultimately aiming at the minimization (and active avoidance) of surprising states (see **Figure 1**). There are several ways of describing the requisite (neuronal) message passing–in terms of Bayesian belief updating (Friston et al., 2017). Perhaps the most popular at present is predictive coding (Rao and Ballard, 1999), where inference and learning is driven by prediction errors, and agency emerges from perception-action loops (Fuster, 2004; Parr and Friston, 2019), continuously exchanging information with the sensorium. By sense of agency we refer to the feeling of control over one's actions and their perceived consequences (Gallagher, 2000; Haggard, 2017).

As underwriting perception and action (Méndez et al., 2014), cognition (i.e., active inference or planning) is closely related to evaluating the consequences of action in relation to prior beliefs about homeostatic needs of survival and reproduction; preparing responses to anticipated change (Pessoa, 2010). Here, beliefs correspond to Bayesian beliefs (i.e., posterior probability distributions over some hidden state of the world)–as opposed to propositional beliefs in the folk psychology sense. Minds and their basic functions–such as perception, emotion, cognition, and action–ultimately seek good predictive control. That is, they are continuously aiming to minimize uncertainty about states of the world, where uncertainty is simply expected surprise (i.e., entropy), given a course of action. There are two fundamental ways to avoid (expected) surprise: (1) change one's cognition, beliefs or hypotheses (i.e., perception), or (2) change the world (i.e., action). This distinction is crucial in the context of robotic systems, which are quintessentially concerned with changing the causes of sensations, rather than changing perceptual inference via cognition (Jovanović et al., 2019).

In short, action aims at reducing uncertainty, where exploratory behavior leads us to interact "freely" with objects in the world–to improve our generative models of the way they behave, maximizing the fit between them, and ultimately rendering these behaviors more predictable (Pisula and Siegel, 2005). A generative model is at the heart of active inference–and indeed the current treatment. Technically, models are a probabilistic specification of how (sensory) consequences are caused by hidden or latent states of the world. It generally comprises a likelihood; namely, the probability of a sensory outcome given a hidden state–and prior beliefs over hidden states. Maximizing the fit or alignment between a generative model of the sensed world–and the process generating sensory outcomes corresponds to minimizing surprise (e.g., prediction error) or–in more statistical terms–maximizing the evidence for their model (Hohwy, 2016). In the setting of active inference, this is often referred to as self-evidencing. In active inference, (expected) surprise is approximated with (expected) variational

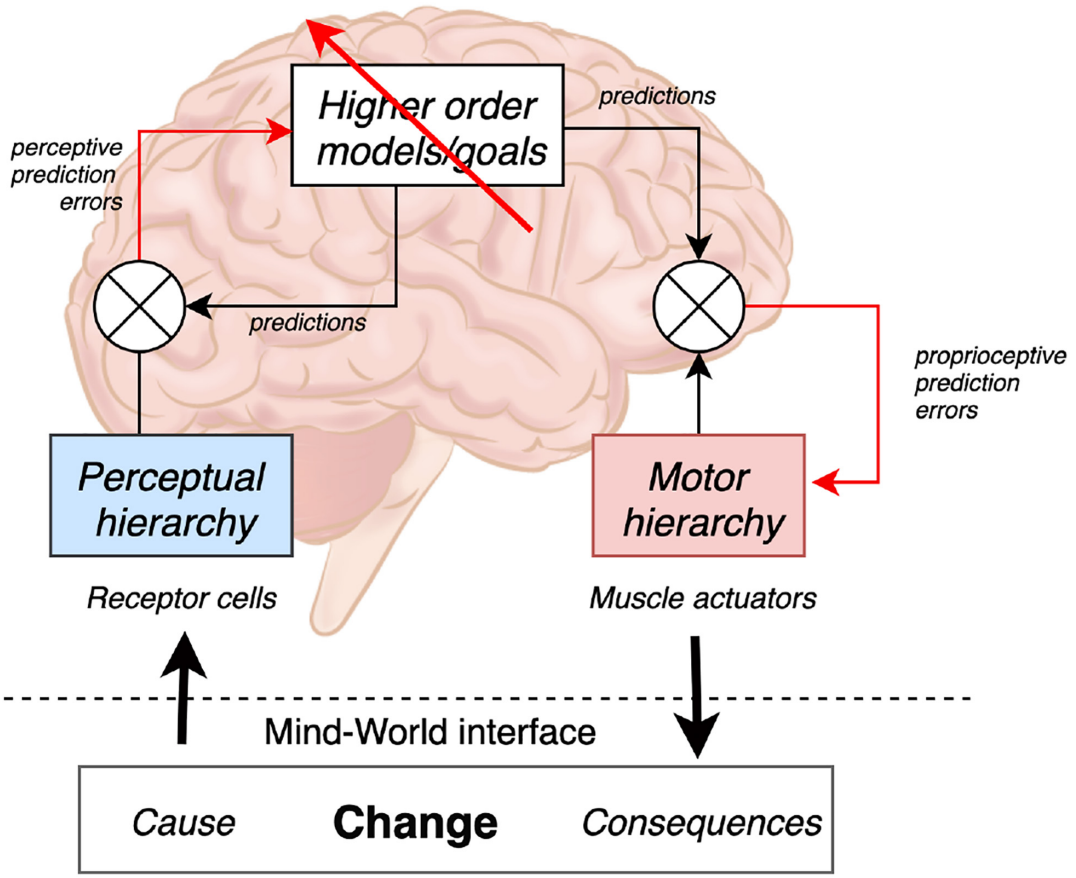


FIGURE 1 | Perception models afferent changes in states of the world detected by receptor cells (e.g., in the retina) all along the perceptual hierarchy. In this control diagram, ⊗ denotes a comparator. The red arrows denote inference and learning (i.e., driven by prediction errors) that compare (descending) predictions with (ascending) sensations. Cognition and higher order processing attempt to predict sensory input and futures states of the world based on available (generative) models; thereby, minimizing prediction error. Action organizes the motor hierarchy in an attempt to actively control the efferent consequences of ongoing events; namely, by modifying causes anticipated through perceptual means, thereby altering the system' dynamics to make them more predictable (i.e., less surprising). Though not specified on this diagram, perception can be further subdivided into interoception and exteroception; respectively, modeling changes in the internal and external world. Emotion–and related notions of selfhood–usually arise via predictive processing of interoceptive sensations, often known as interoceptive inference (Seth, 2013, 2014; Seth and Friston, 2016).

free energy; thereby providing a tractable objective function for perception and action. The integration of efferent (motor) and afferent (sensory) signals results in what can be termed the sensation of control, or feeling of agency, whereby sensorimotor mismatch is minimized.

These three functions of perception-cognition-action form a hierarchical system with sensorimotor signals at the lowest levels of the hierarchy, and abstract cognition (executive functions of goal- directed planning and decision-making) at the highest levels (Schoeller et al., 2018). Perception is organized in a hierarchical fashion, with bottom-up sensory signals (e.g., "a change in color from red to green") being continuously predicted by top-down cognitive models (e.g., "green-light authorization for crossing the street"). Action models are also organized hierarchically, whereby fine motor interaction with the external world (e.g., typing on a keyboard), are contextualized by higher order goals (e.g., writing a paragraph), themselves prescribed by high abstract plans (e.g., getting a paper accepted in a conference)–ultimately underwriting existential goals–corresponding to the organization of life itself (Schoeller et al., 2018).

A key notion is precision weighting, which refers to the reliability or salience of prediction errors. The higher the precision, the more impactful the prediction errors on how processing unfolds. In Active Inference terms, precision represents the agent's confidence that certain action policies (i.e., sequence of actions) will produce the states the agent highly expects (Friston et al., 2014). Predictive agents decide what actions to pursue based on the predicted sensory consequences of the action–choosing those behaviors that are most likely to minimize surprise over the long term, and so maximize their time spent in the sensory states they expect. The performance of action policies to reduce prediction error can be plotted as a slope that depicts the speed at which errors are predicted to be managed along the way. The steepness of the slope indicated how fast errors are being reduced given some policy: the steeper the slope

the faster the rate, the shallower the slope the slower the rate. If the speed of error reduction is faster that expected, the action policy should be made more precise; and if the rate is slower than expended, and errors are amassing unexpectedly, then the policy isn't as successful at bringing about those future sensory states that are expected, and this should be taken as evidence for weighing an action policy as having low precision.

Change in the rate at which error is being resolved manifests for humans as emotional valence–we feel good when error is being reduced at a better than expected rate, and we feel bad when error is unexpectedly on the rise (Joffily and Coricelli, 2013; Schoeller, 2015, 2017; Schoeller and Perlovsky, 2016; Schoeller et al., 2017; Van de Cruys, 2017; Kiverstein et al., 2019; Perlovsky and Schoeller, 2019; Wilkinson et al., 2019; Nave et al., 2020). Valence systems provide the agent with a domain general controller capable of tracking changes in error managements and adjusting precision expectations relative to those changes (Kiverstein et al., 2019; Hesp et al., 2021). This bodily information is a reflection of an agent's perceived fitness–that is, how adaptive the agent's current predictive model is relative to their environment.

Affective valence is widely acknowledged to play an important role in trust (Dunn and Schweitzer, 2005). Positive feelings have been shown to increase trusting, while negative feelings diminish it (Dunn and Schweitzer, 2005). The active inference framework helps to account for this evidence, suggesting that positive and negative feelings are in part a reflection of how well or poorly one is able to predict the actions of another person. As detailed in the following section, affectivity plays a crucial role in mediating exchanges with robots, often acting as a cardinal determinant of trust in that context specifically (Broadbent et al., 2007). As a consequence, robotic design that considers affect–and related higher-level constructs–are likely to enhance productivity and acceptance (Norman et al., 2003).

AGENCY AND EMPOWERMENT IN HUMAN-TECHNOLOGICAL EXTENSION

The relevance of active inference for robotics has been experimentally demonstrated in Pio-Lopez et al. (2016). In the context of automation, understanding human agency is all the more important–as experimental studies have demonstrated that one can prime for agency with external cues (leading to abusive control), and clinical studies reveal that an impairment of control is associated with depression, stress, and anxiety-related disorders (Abramson et al., 1989; Chorpita and Barlow, 1998). The integration of efferent (motor) and afferent (sensory) signals results in what can be termed the sensation of control or a feeling of agency (Salomon et al., 2016; Vuorre and Metcalfe, 2016), which depends on the correspondence of top-down (virtual) predictions of the outcomes of action, and the bottom-up (actual) sensations. As illustrated in **Figure 1**, the brain compares actual sensory consequences of the motor action with an internal model of its predicted sensory consequences. When predicted sensory consequences match incoming sensory signals, the movement is attributed to the self and a (confident) sense of agency is said to emerge (Wolpert et al., 1995; Hohwy, 2007; Synofzik et al., 2008; Salomon et al., 2016). Situations where there is a mismatch between intended and observed actions we also see a feeling of loss of agency, and an attribution of the movement (or lack thereof) to an external source. For example, if someone was to move my arm then there would be the sensory experience but without the prediction. If instead I was to try to move my arm, but due to anesthetic I was unable to, there would be the prediction but not the sensory confirmation. Agency then is just another hypothesis (or Bayesian belief) that is used to explain interoceptive, exteroceptive, and proprioceptive input. If sensory evidence is consistent with my motor plans, then I can be confident that "I caused that." Conversely, if I sense something that I did not predict, then the alternative hypothesis that "you caused that" becomes the best explanation (Seth, 2015). The accompanying uncertainty may be associated with negative affect such as stress or anxiety (Stephan et al., 2016; Peters et al., 2017). Again, the very notions of stress and anxiety are treated as higher-level constructs–that best explain the interoceptive signals that attend situations of uncertainty and adjust precision accordingly; e.g., physiological autonomic responses of the flight or fright sort (Barrett and Simmons, 2015; Seth and Friston, 2016).

To measure the amount of control (or influence) an agent has and perceives, Klyubin et al. (2005) proposed the concept of empowerment. Empowerment is a property of self-organized adaptive systems and is a function of the agent perception-action loop, more specifically the relation between sensors and actuators of the organism, as induced by interactions between the environment and the agent's morphology (Salge and Polani, 2017). Empowerment is low when the agent has no control over what it senses, and it is high the more control is evinced (Friston et al., 2006). An information-theoretic definition has been proposed, whereby empowerment is interpreted as the amount of information the agent can exchange with its environment through its perception-action cycle. According to Klyubin et al. (2005), empowerment is null when the agent has no control over what it is sensing, and it is higher the more perceivable control or influence the agent has. Hence, "empowerment can be interpreted as the amount of information the agent could potentially inject into the environment via its actuator and later capture via its sensor." Consider for example the difference between passively watching a movie and being engaged with the same content in an immersive virtual reality setting. Crucially, empowerment is a reflection of what an agent *can* do, not what the agent actually does (Klyubin et al., 2005), and maximizing empowerment adapts sensors and actuators to each other. In other words, empowerment can be described in terms of sensorimotor fitness–i.e., the spatial and temporal relevance of the feedback the robot gets on its own behavior. For example, a robot that gets multisensor feedback on the probability of success of its actions has greater empowerment than a robot who is deprived of, say, visual information or which receives delayed information (the greater the delay, the weaker the empowerment). This calls forth a framework where the so-called exploration/exploitation dilemma (crucial for safety in HRC) can be casted as a behavioral account of the perception-action cycle.

Technology considerably increases human empowerment (Brey, 2000), freeing the human animal from many niches or geographical constraints (e.g., climate or geology), and allowing increasingly complicated narratives and trajectories to develop within the scope of human control (e.g., cranes allow the manipulation of heavy systems beyond mere human capabilities). Predictive organisms are attracted to–and rewarded by–opportunities to improve their predictive grip on their environments–i.e., to improve their empowerment. By definition, technological extension of the perception-action cycle offers a powerful way of expanding empowerment, but to function effectively it needs to be integrated with the agent's sensorimotor dynamics. In other words, technology must enter the agent' extended repertoire of behaviors. That inclusion requires the technological extensions to be modeled internally by the agent in the same capacity of its own sensorimotor contingencies, at some level of abstraction. This (self) modeling of technological extension is key to the emergence of trust–in active inference terms: a high precision on beliefs about how the technology will behave and evolve relative to our own sensorimotor engagements. This is an extension of the same mechanism giving rise to agency beyond the realm of the body. As we attempt to show in the next section, this extension of human control beyond mere motor action and its cognitive monitoring requires trust–as a sense of virtual control in an extended perception-action cycle (Sheridan, 1988). The study of human agency has clear relevance for robotic motor control, but to our knowledge it has not yet been applied to the problem of trust in complex technological systems or human-robot interaction. In the next section, we examine the possibility of modeling trust in relation to active inference and empowerment.

TRUST AS VIRTUAL CONTROL IN EXTENDED AGENCY

Within the context of human-robot interactions (Lee, 2008), optimal trust is crucial to avoid so-called disuse of technology (i.e., loss of productivity resulting from users not trusting the system), but also abuse of technology (i.e., loss of safety resulting from overreliance on the system). Hence, the cognitive neuroscience of trust has implications for both safety and management (Sheridan and Parasuraman, 2005; Lee, 2008). Indeed, technological abuse and overreliance on automation count among the most important sources of catastrophes (Sheridan and Parasuraman, 2005). From a theoretical point of view, tremendous variations exist in what trust represents and how it can best be quantified, and several definitions have been suggested with potential applications for automation (Muir, 1994; Cohen et al., 1999). An exhaustive review–of the large body of work devoted to trust literature–is outside the scope of this article: excellent reviews can be found in Lee and See (2004) and Sheridan (2019b). Here, we present the fundamental elements of these models of trust, in the light of perception-action loops, and potential applications to robotics to demonstrate the relevance of the active inference framework for human factors in HRI.

Several measures of trust exist in a variety of settings from management, to interpersonal, and automation. In reviewing the literature on trust, Lee and See identified three categories of definitions; all fundamentally related to uncertainty and control (2004). The fundamental relation between trust and uncertainty appears most salient in situations when the uncertainty derives from the realization of goals or intentions (e.g., in human-robot interactions, or employee-employer relationships), where internal details about the agent are unknown, leaving the trustor vulnerable. In the context of robotics–where human action is extended by robotic systems–the match between goals of the (extended) human agent and those of the (extending) robotic agent is crucial in determining the success of the relation (whether the agent will make use of the extension). In order of generality, the definitions identified by Lee and See are: (1) trust as intention to (contract) vulnerability, (2) trust as vulnerability, and (3) trust as estimation of an event likelihood. Note that these three general definitions, derive from early definitions of trust by Muir (1994) and Mayer et al. (1995), according to whom trust is a trade-off between ability (A) and benevolence (B), whereby a reliable system is high in both A and B (**Figure 2**).

The importance of externalizing goals of robotics systems (i.e., transparency) at all levels of the hierarchical perception-action loop cannot be stressed enough–for successful communication and gradual building of trust (Sheridan and Parasuraman, 2005). This is well captured in the standard definition of trust by Sheridan (2019b), where communication of goals (or transparency) plays a crucial role among the seven item scales of trust (see **Table 1**).

In summary, trust is fundamentally related to human control to the extent that it is required for any extension of the perception-action cycle (i.e., when the success of the performance depends on some other agent's perception-action cycle, rather than one's own). Above, we saw that vulnerability is a function of empowerment in the extended agent (the more extended the agent, the more vulnerable), which can be evaluated through interaction with the robotic perception-action cycle. This may help to explain why operator curiosity is an important source of accidents in the robot industry (Lind, 2009), as curiosity aims to reduce uncertainty about the technology and so increase trust and control, and suggests potential solutions in the field of

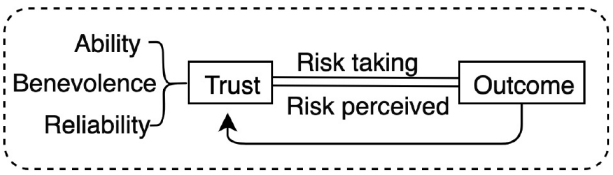


FIGURE 2 | Muir and Mayer model of trust as a function of the trustee's ability, benevolence and reliability (1995) where risk perception affects risk action. This bipartition of trust as ability and benevolence amounts to two different levels in the motor hierarchy of the extended agent (e.g., the robot), whereby benevolence refers to the high-level goals motivating the extended agent and ability refers to the means of the agent to realize these goals, i.e., the sophistication of its low-level motor output in relation to the task at hand.

TABLE 1 | Standard definition of trust by Sheridan (2019b).

(1) Statistical reliability (lack of error).
(2) Usefulness (ability of the system to do what is most important, e.g., in trading benefits and costs).
(3) Robustness (ability and flexibility of the system to perform variations of the task).
(4) Understandability (transparency of the system in revealing how and why it is doing what it is doing).
(5) Explication of intent (system communicating to the trustee what it will do next).
(6) Familiarity (to the user based on past experience).
(7) Dependence (upon the system by the trustee as compared to other ways of doing the given task).

accidentology. Trust is required in situations of uncertainty; and it varies as the system exhibits predictable regularities. Sheridan and Meyer models suggest that one will trust a predictable system, to the extent that one can act upon that system to obtain similar results over time, and eventually render its behavior more predictable through incremental alterations.

We have considered how a sense of agency emerges, as the resolution of mismatch between (1) the (perceptual) expectation (i.e., hypothesis) about the consequences of (motor) action, and (2) the perceived results of action (observation, perception). We introduced the idea of trust as a sense of virtual, extended control. In other words, trust is a measure of the precision, or confidence, afforded by action plans that involve another (i.e., of the match between one's actions–and their underlying intentions–and the predicted sensory consequences through another agent). As such, "trust" is an essential inference about states of affairs; in which the anticipated consequences of extended action are realized reliably. From the point of view of "emotional" inference (Smith et al., 2019), trust is therefore the best explanation for a reliable sensory exchange with an extended motor plant or partner. Given the role that affect plays in tuning precision on action policies, "reliable" here means a reliable way to reduce expected free energy (via the extended interaction). We are attracted by, or solicited to use, a tool or device because it affords to us a means of reducing error, in a better than expected way relative to doing the same work in the absence of technological extension.

It is generally assumed that trust in any system increases with evidence of that system's reliability (**Figure 3**). The greater the convergence of behavior models between trustor and trustee (i.e., the largest the benevolence), the greater the trust in the relationship (Hisnanick, 1989). Perhaps, this explains why simple mimicry facilitates adoption, or why one tends to agree with people who behave similarly–we generalize shared goals on the basis of shared behavior (Cirelli, 2018). The similarity-attraction hypothesis in social psychology predicts that people with similar personality characteristics will be attracted to each other (Morry, 2005). Hence, technology that displays personality characteristics–similar to those of the user–tends to be accepted more rapidly (Nass et al., 1995). As machines become increasingly intelligent, it is crucial that they communicate higher-order goals accordingly (Sheridan, 2019b). Communication of goals can be simplified by rendering the perception-action cycle explicit/and augmenting sensors to indicate their perceptual range (e.g., the human retina affords some information about the portion of the visual field it senses); thereby, greatly reducing the risk of accidents.

Finally, trust is a fundamentally dynamic process that eventually leads to a state of dependence (**Figure 4**). This is best exemplified in the context of information technology, whereby the information is no longer stored internally (e.g., phone numbers, navigation pathways, historical facts) but all that is known is the access pathway (my phone's contact list, my preferred web mapping service, a Wikipedia page). As suggested by the Sheridan scale, the dynamics of trust go beyond mere predictability and ultimately lead to a state of prosthetic dependence in the context of the specific task. This is evident in the context of automation, which increases the perception-action cycle at an exponential rate, thereby leading to a high abandon rate of past practices, as new technologies are adopted. Formally speaking, as technology allows the agent to reduce prediction error (by better understanding the problem space, and through more empowered actions) the agent comes to expect that slope of error reduction within those contexts and relative to the specific tasks. The result is a gradual loss of interest or solicitation by previous less potent forms of HRCs–they have become outdated and so have lost their motivational appeal.

In the context of interpersonal relationships, Rempel et al. (1985) described trust as an evolving phenomenon, where growth is a function of the relationships progress. They further argue that the anticipation of future behavior forms the basis of trust at the earliest stages of a relationship. This is followed by dependability, which reflects the degree to which behavior is consistent. As the relationship matures, the basis of trust ultimately passes the threshold of faith, which has been related to benevolence (Lee and See, 2004); i.e., coordination on higher order goals driving behavior. Crucially, an early study of the adaptation of operators to new technology demonstrated a similar progression (Hisnanick, 1989). Trust in that context depends on trial-and-error experience, followed by understanding of the technology's operation, and finally, a state of certainty or faith (see **Figure 5**). Lee and Moray (1992) made similar distinctions in defining the factors that influence trust in automation.

TRUST DURING DYADIC COLLABORATION

We have seen that the essential components of trust (benevolence and competence) can be cast in terms of the confidence in beliefs at (respectively) high and low levels in the motor hierarchy, but how can active inference contribute to the science of extended agency? In this section, we examine the role of

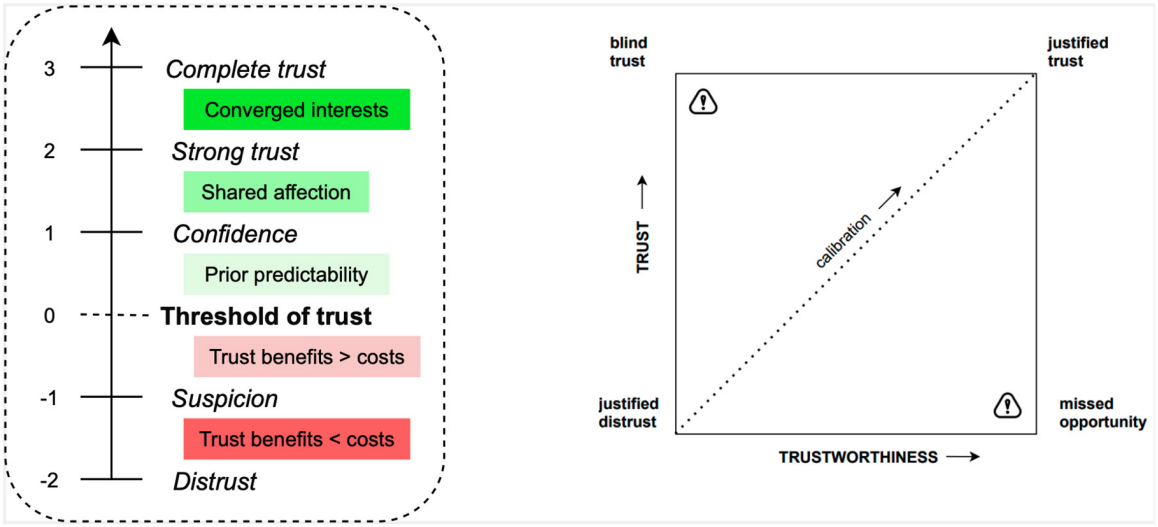


FIGURE 3 | On the left, levels of trust from Dietz and Den Hartog (2006). On the right, cross-plot of (objective) trustworthiness compared to (subjective) trust by Sheridan (1988), Sheridan (2019b). As a pioneer in the study of trust in technology, Sheridan further suggested that (subjective) trust can be cross-plotted against (objective) trustworthiness. This representation engenders four extremes: justified trust or distrust, blind trust (trusted untrustworthy; i.e., misuse) and missed opportunity (untrusted trustworthy; i.e., disuse). The dotted curve represents calibration, which is linear when trust is justified. Poor calibration can lead to loss of safety (due to overconfident misuse), or loss of productivity (due to underconfident disuse).

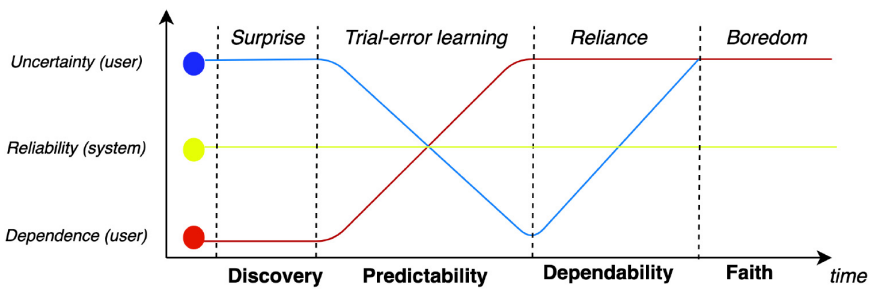


FIGURE 4 | Dynamics of trust over time–with four phases from discovery to faith: for a consistently reliable system, dependence (i.e., risk) is inversely proportional to uncertainty, assessed through a cycle of trial and error, until a threshold is reached. Through cycles of trial and errors, trust gradually evolves from predictability (model) to dependability (control) to a state of faith (overreliance). Our model suggests that boredom is a marker of overreliance.

expectations in the context of dyadic interaction. So, what would a formal (first principles) approach like active inference bring to HRC? At its most straightforward, trust is a measure of the confidence that we place in something behaving in beneficial ways that we can highly predict. Technically, this speaks to the encoding of uncertainty in generative models of dyadic interactions. These generative models necessarily entail making inferences about policies; namely, ordered sequences of action during dyadic exchanges (Moutoussis et al., 2014; Friston and Frith, 2015). This could range from turn taking in communication (Wilson and Wilson, 2005; Ghazanfar and Takahashi, 2014) to skilled interactions with robotic devices. At its most elemental, the encoding of uncertainty in generative models is usually framed in terms of the precision (i.e., inverse variance) or confidence (Friston et al., 2014). Crucially, every (subpersonal) belief that is updated during active inference can have the attribute of a precision or confidence. This means that the questions about trust reduce to identifying what kind of belief structure has a precision that can be associated with the construct of "trust." In generative models based upon discrete-state spaces (e.g., partially observed Markov decision processes) there are several candidates for such beliefs. Perhaps the most pertinent–to dyadic interactions–are the beliefs about state transitions; i.e., what happens if I (or you) do that. For example, if I trust you, that means I have precise Bayesian beliefs about how you will respond to my actions. This translates into precise beliefs about state transitions during controlled exchanges (Parr and Friston, 2017; Parr et al., 2018). This means that I can plan deep into the future before things become very uncertain and, in turn, form precise posterior beliefs about the best courses of action, in other words our policies align (see **Figure 5**).

Conversely, if I do not trust you, I will have imprecise beliefs about how you will respond and will only be able to entertain

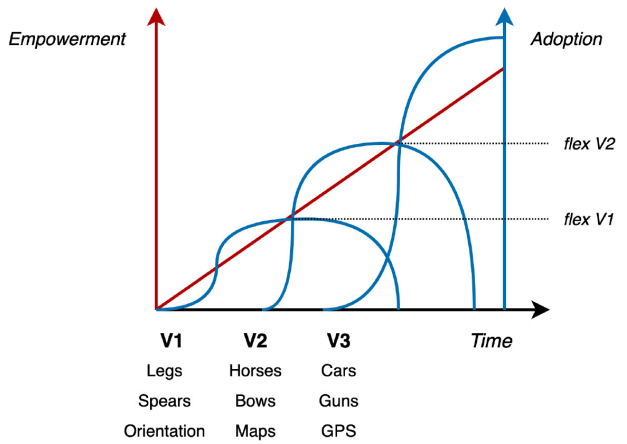


FIGURE 5 | Waves of technological adoption related to predictive slopes of extended engagement (empowerment) during versioning of the technology. Indeed this is an oversimplification for the sake of visualization as we are assuming a linear progression of empowerment over time in the evolving versions of the technology (i.e., a healthy research and development cycle) where, for most technologies, newer versions may not present much greater empowerment as compared to older ones. The important idea here is the inflection point (flex) indicating the start of technological decay reflecting the abandon rate of a practice as the experience of better predictive slopes of extended technological engagements lead to disengagement of non-extended approaches (e.g., cars replace horses replacing legs). Old slopes are less than expected and so unsatisfactory as compared to new ones.

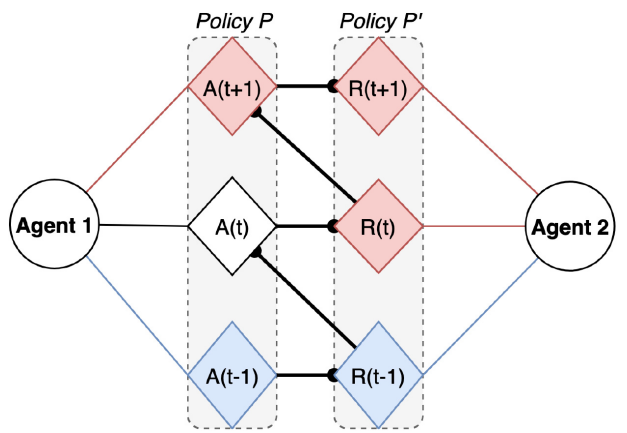


FIGURE 6 | A trust dyad, whereby Agent 1 performs action A at t, within the action policy P, and in collaboration with Agent 2. In a trustworthy relationship, Agent 1 can expect from Agent 2 an action Policy P', where P' is symmetrical to P (each action of P' at t + 1 is a response to P at t). Past (observed) actions are blue and future (anticipated) actions are red. The bold line in between policies represents the shared policy (or joint narrative), whereby A(t + 1) can be prepared based on beliefs about anticipated R(t).

short term plans during any exchange. Furthermore, it will be difficult to infer precise outcomes of any course of action–and hence hard to entertain a shared policy. This means I will also be uncertain about which is the best course of action. Technically, this results in an imprecise belief distribution over policies, which is normally associated with negative affect or some form of angst (Seth and Friston, 2016; Badcock et al., 2017; Peters et al., 2017). Notice, that now there is not just error in the environment to deal with but also the uncertainty of the other. As uncertainty increases, negatively valenced feelings emerge as a reflection of that change, and in turn reduce precision on the policies related to that collaboration. The result is the agent is less likely to be attracted to enact policies of extension with that other person or robot, and so much more likely to revert to using more habitual (and already highly precise) ways of reducing error. In short, almost by definition, engaging with an untrustworthy partner is, in a folk psychological sense, rather stressful.

Clearly, this active inference formulation is somewhat hypothetical. There will be many other belief structures that could be imprecise; for example, prior beliefs about the policies I should entertain and, indeed, the precision of likelihood mappings (that map from latent or hidden states of the world to observed outcomes). The latter is usually considered in terms of ambiguity (Friston et al., 2017; Veissière et al., 2019). In other words, I could consider your behavior or responses ambiguous–and that could render you untrustworthy; even if I have very precise beliefs about the latent states you are likely to navigate or pursue. In short, it may be an open question as to whether the precision of state transitions, likelihood contingencies or prior beliefs about policies manifest as differences in trust. This brings us to a fundamental motivation for the formalization of trust in terms of active inference.

It is possible to build models of dyadic exchange under ideal Bayesian assumptions using active inference (e.g., Moutoussis et al., 2014; Friston and Frith, 2015). This means that one can optimize the prior beliefs inherent in these models to render observed choice behavior the most likely. Put another way, one can fit active inference models to empirical behavior to estimate the prior beliefs that different subjects evince through their responses (Parr et al., 2018). These estimates include a subject's prior beliefs about the precision of various probability distributions or Bayesian beliefs. In turn, this means it should be possible to phenotype any given person in an experimentally controlled (dyadic) situation and estimate the precision of various beliefs that best explain their behavior. One could, in principle, then establish correlations between different kinds of precision and other validated measures of trust, such as those above. This would then establish what part of active inference best corresponds to the folk psychological–and formal definitions of trust. Interestingly, this kind of approach has already been considered in the context of computational psychiatry and computational phenotyping; especially in relation to epistemic trust (Fonagy and Allison, 2014). Epistemic trust is a characteristic of the confidence placed in someone as a source of knowledge or guidance. Clearly, this kind of trust becomes essential in terms of therapeutic relationships and, perhaps, teacher pupil relationships. Finally, one important determinant of the confidence placed in–or precision afforded–generative models of interpersonal exchange is the degree to which I can use myself as a model of you. This speaks to the fundamental importance of a shared narrative (or generative model) that underwrites any meaningful interaction of the sort we are talking about. This can be

articulated in terms of a generalized synchrony that enables a primitive form of communication or hermeneutics (Friston and Frith, 2015). Crucially, two agents adopting the same model can predict each other's behavior, and minimize their mutual prediction errors (**Figure 6**). This has important experimental implications, especially in the context of HRC, where robotic mimicry can be seen as mere self-extension for the user, leading to what philosophers of technology call relative transparency (where whatever impacts the robot also impacts me–see Brey, 2000). The self being the product of the highest prediction capacities, when another agent becomes more predictable it also increases the similarity at the highest levels in the cognitive hierarchy and thereby facilitates joint action.

This mutual predictability is also self-evident in terms of sharing the same narrative; e.g., language. In other words, my modeling of you is licensed as precise or trustworthy if, and only if, we speak the same language. This perspective can be unpacked in many directions; for example, in terms of niche construction and communication among multiple conspecifics (in an ecological context) (Constant et al., 2019; Veissière et al., 2019). It also speaks to the potential importance of taking into account self-models in HRC design, allowing both users and robots to represent each other's behavior efficiently. Indeed, on the above reading of active inference, such shared narratives become imperative for trustworthy exchanges and collaboration. Indeed, current models suggest that the rise of subjectivity and the "self" are grounded in privileged predictive capacities regarding the states of the organism compared to the external environment (Limanowski and Blankenburg, 2013; Apps and Tsakiris, 2014; Allen and Friston, 2016; Salomon, 2017). As such, dyadic trust in another agent (biological or artificial) can be viewed as a process of extending these predictive processes beyond the body and rendering the external agent as part of a self model. Moreover, recently robotic interfaces have been used to induce modulations of self models by interfering with sensorimotor predictions. This in turn gives rise to phenomena closely resembling psychiatric symptoms (Blanke et al., 2014; Faivre et al., 2020; Salomon et al., 2020).

CONCLUSION

In the light of our increasing dependence on technology, it is worth considering that the largest aspect of human interactions with machines (their use) essentially rests upon vague approximative mental models of the underlying mechanisms (e.g., few smartphone users can understand the functioning of a computer operating software). Technically, in active inference, the use of simplified generative models (e.g., heuristics) is an integral part of self-evidencing. This follows because the evidence for a generative model (e.g., of how a smartphone works) can be expressed as accuracy minus complexity. In this setting, complexity is the divergence between posterior and prior beliefs–before and after belief updating. This means the generative model is required to provide an accurate account of sensory exchanges (with a smartphone) that is as simple as possible (Maisto et al., 2015). In short, the best generative model will be, necessarily, simpler than the thing it is modeling. This principle holds true of technology in general (extending the scope of human perception-action cycles), and automation specifically (replacing these perception-action capabilities). We have examined the concept of trust from the standpoint of control and perception-action loops and found that trust components (i.e., competence and benevolence) are best casted in terms of an action-cognitive hierarchy. By examining trust from the standpoint of active inference, we were also better able to understand phenomena, such as exploration-related accidents, and the gradual building of trust with shared goals, narratives and agency. One of the benefits of this model is that it applies to any sort of collaborative enterprise between humans and machines. Although the specifications of the machine (e.g., its size, its use, etc.) and the nature of the collaboration (e.g., occasional, constant, autonomous, etc.) will of course change how and what one models about the collaborative machine, the trust one feel emerges from the identical process of modeling their states and behaviors over time in ways that allow them to be included in one's own generative model (in a particular context). HRC is of course only a first step and it will be interesting going forward to consider how this model of trust as extended predictive control practically is applied to the wide variety of cases where humans and machines are working closely together in our world today.

As the complexity and autonomy of artificial systems go up, so too will the complexity and sophistication of the model we generate about the behaviors of those systems. In the case of collaborating with artificial intelligence systems this becomes even more challenging, and would increasingly require useful opacifications of the underlying decision making mechanisms that drive those system's behaviors. The science of human-robot interaction could make rapid progress if objective measures of trust were developed, and the neuroscience of agency does offer such metrics. It is here that a simulation setup of the sort offered by active inference could play an important part. Among the potential biomarkers for agency and control, the N1 component of event related electrical brain responses–a negative potential occurring approximately 100 ms after stimulus onset–is attenuated during self-produced or predicted events, relative to that observed during externally generated feedback. As machine become increasingly intelligent, it is to be expected that not only users will develop more sophisticated (generative) models of their internal behavior and the reliability of these behavior, but robots will also adapt to interindividual differences (Sheridan, 2019b), hence reciprocally monitor the trustworthiness of users, and thereby allow for safer and more productive interaction.

In this paper we have proposed a novel view of trust as extended (predictive) control, a view that is well poised to help us elucidate the mechanisms underlying trust between humans, and between humans and technological artifacts. However, this should only be seen as the beginning. The field of HRC is quickly

evolving, as the robots we find ourselves collaborating with are increasingly complex and autonomous. Degree of autonomy is of particular importance here for thinking about HRCs. As autonomy increases in our robotic partners different forms of collaboration are bound to emerge, and new requirements for trusting those artifacts will be necessary. While we do not have the space here to fully explore these more complex examples in current and future HRC, we can at least say that transparency and ethical-design will become increasingly important. Given the framework we have proposed, for trust to emerge in these complex interactions human agents need to be able to accurately (or at least usefully) model the sorts of decision-trees that the autonomous artificial agents make use of in various contexts. The means by which such transparency can be achieved is a topic for further research.

Furthermore, as artificial intelligence systems evolve in complexity we will inevitably be interacting with technological artifacts that are able to model humans in return. This two-way predictive modeling will result in new forms of collaboration and new approaches to developing a trusting relationship (see Demekas et al., 2020). Collaborative dynamics between humans is already being modeled using the AIF (Ramstead et al., 2020), in which predictive agents model each other's generative model in ways that allow groups to temporarily become a unified error-minimizing machine. With the possibility of future artificial autonomous agents using variations of a prediction hierarchy like humans use, exploring the emergent dynamics between human and artificial agents in this way becomes possible as well.

KEY POINTS:

- Mind–all brain–is a constructive, statistical organ that continuously generates hypotheses to predict the most likely causes of its sensory data.
- We present a model of trust as the best explanation for a reliable sensory exchange with an extended motor plant or partner.
- User boredom may be a marker of overreliance.
- Shared narratives, mutual predictability, and self-models are crucial in human-robot interaction design and imperative for trustworthy exchanges and collaboration.
- Generalized synchrony enables a primitive form of communication.
- Shared generative models may allow agents to predict each other more accurately and minimize their prediction errors or surprise, leading to more efficient HRC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

MM carried out this work with the support of Horizon 2020 European Union ERC Advanced Grant XSPECT – DLV-692739.

REFERENCES

Abramson, L. Y., Metalsky, G. I., and Alloy, L. B. (1989). Hopelessness depression: a theory-based subtype of depression. *Psychol. Rev.* 96, 358–372. doi: 10.1037/0033-295x.96.2.358

Allen, M., and Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482. doi: 10.1007/s11229-016-1288-5

Apps, M. A., and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* 41, 85–97. doi: 10.1016/j.neubiorev.2013.01.029

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., and Friston, K. J. (2017). The depressed brain: an evolutionary systems theory. *Trends Cogn. Sci.* 21, 182–194. doi: 10.1016/j.tics.2017.01.005

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Blanke, O., Pozeg, P., Hara, M., Heydrich, L., Serino, A., Yamamoto, A., et al. (2014). Neurological and robot-controlled induction of an apparition. *Curr. Biol.* 24, 2681–2686. doi: 10.1016/j.cub.2014.09.049

Brey, P. (2000). "Technology as extension of human faculties," in *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology*, ed. C. Mitcham (London: Elsevier/JAI Press).

Broadbent, E., MacDonald, B., Jago, L., Juergens, M., and Mazharullah, O. (2007). "Human reactions to good and bad robots," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. Presented at the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Piscataway, NJ: IEEE).

Chorpita, B. F., and Barlow, D. H. (1998). The development of anxiety: the role of control in the early environment. *Psychol. Bull.* 124, 3–21. doi: 10.1037/0033-2909.124.1.3

Cirelli, L. K. (2018). How interpersonal synchrony facilitates early prosocial behavior. *Curr. Opin. Psychol.* 20, 35–39. doi: 10.1016/j.copsyc.2017.08.009

Cohen, M. S., Parasuraman, R., and Freeman, J. T. (1999). *Trust in Decision Aids: a Model and its Training Implications*. Arlington, VA: Cognitive Technologies. Technical Report USAATCOM TR 97-D-4.

Constant, A., Ramstead, M. J., Veissière, S. P., and Friston, K. (2019). Regimes of expectations: an active inference model of social conformity and human decision making. *Front. Psychol.* 10:679. doi: 10.3389/fpsyg.2019.00679

Demekas, D., Parr, T., and Friston, K. J. (2020). An investigation of the free energy principle for emotion recognition. *Front. Comp. Neurosci.* 14:30. doi: 10.3389/fncom.2020.00030

Dietz, G., and Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Rev.* 35, 557–588. doi: 10.1108/00483480610682299

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). "Explainable artificial intelligence: a survey," in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (Piscataway, NJ: IEEE), 0210–0215.

Dunn, J. R., and Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *J. Pers. Soc. Psychol.* 88, 736–748. doi: 10.1037/0022-3514.88.5.736

Faivre, N., Vuillaume, L., Bernasconi, F., Salomon, R., Blanke, O., and Cleeremans, A. (2020). Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex* 124, 224–234. doi: 10.1016/j.cortex.2019.12.001

Fonagy, P., and Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy* 51:372. doi: 10.1037/a0036505

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain (archive). *J. Physiol. Paris* 100, 70–87.

Friston, K. J., Parr, T., and de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neurosci. (Cambridge, Mass)* 1, 381–414. doi: 10.1162/netn_a_00018

Friston, K. J., Stephan, K. E., Montague, R., and Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry* 1, 148–158. doi: 10.1016/S2215-0366(14)70275-70275

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends Cogn. Sci.* 8, 143–145. doi: 10.1016/j.tics.2004.02.004

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/s1364-6613(99)01417-5

Ghazanfar, A. A., and Takahashi, D. Y. (2014). The evolution of speech: vision, rhythm, cooperation. *Trends Cogn. Sci.* 18, 543–553. doi: 10.1016/j.tics.2014.06.004

Gregory, R. L. (1980). Perceptions as hypotheses (archive). *Phil. Trans. R. Soc. Lond. B* 290, 181–197.

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 197–208.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput.* 33, 398–446. doi: 10.1162/neco_a_01341

Hisnanick, J. (1989). In the age of the smart machine: the future of work and power. *Emp. Respons. Rights J.* 2, 313–314.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche* 13, 1–20.

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comp. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Jovanović, K., Petrič, T., Tsuji, T., and Oddo, C. M. (2019). Editorial: human-like advances in robotics: motion, actuation, sensing, cognition and control. *Front. Neurorobot.* 13:85. doi: 10.3389/fnbot.2019.00085

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*, 4th Edn. New York City, NY: McGraw Hill Companies.

Kiverstein, J., Miller, M., and Rietveld, E. (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese* 196, 2847–2869. doi: 10.1007/s11229-017-1583-9

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). "Empowerment: a universal agent-centric measure of control," in *Proceedings of the Congress on Evolutionary Computation*, (Piscataway, NJ: IEEE), 128–135.

Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J., and See, K. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lee, J. D. (2008). Review of a pivotal human factors article: "humans and automation: use, misuse, disuse, abuse.". *Hum. Factors J. Hum. Factors Ergonom. Soc.* 50, 404–410. doi: 10.1518/001872008x288547

Limanowski, J., and Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Front. Hum. Neurosci.* 7:547. doi: 10.3389/fnhum.2013.00547

Lind, S. (2009). *Accident Sources in Industrial Maintenance Operations. Proposals for Identification, Modelling and Management of Accident Risks (Tapaturmat teollisuuden kunnossapitotöissä - Ehdotuksia tapaturmariskien tunnistamiseen, mallinnukseen ja hallintaan)*. Espoo: VTT Publications.

Maisto, D., Donnarumma, F., and Pezzulo, G. (2015). Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving. *J. R. Soc. Interface* 12:20141335. doi: 10.1098/rsif.2014.1335

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

Méndez, J. C., Pérez, O., Prado, L., and Merchant, H. (2014). Linking perception, cognition, and action: psychophysical observations and neural network modelling. *PLoS One* 9:e102553. doi: 10.1371/journal.pone.0102553

Morry, M. M. (2005). Relationship satisfaction as a predictor of similarity ratings: a test of the attraction-similarity hypothesis. *J. Soc. Personal Relationships* 22, 561–584. doi: 10.1177/0265407505054524

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. *Conscious Cogn.* 25, 67–76. doi: 10.1016/j.concog.2014.01.009

Muir, B. M. (1994). Trust in automation: Part I. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 1905–1922. doi: 10.1080/00140139408964957

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities? *Int. J. Human-Computer Stud.* 43, 223–239. doi: 10.1006/ijhc.1995.1042

Nave, K., Deane, G., Miller, M., and Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Rev. Cogn. Sci.* 11:e1542.

Norman, D. A., Ortony, A., and Russell, D. M. (2003). Affect and machine design: lessons for the development of autonomous machines. *IBM Systems J.* 42, 38–44. doi: 10.1147/sj.421.0038

Parr, T., and Friston, K. J. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14:20170376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. J. (2019). Generalised free energy and active inference. *Biol. Cybern.* 113, 495–513. doi: 10.1007/s00422-019-00805-w

Parr, T., Rees, G., and Friston, K. J. (2018). Computational neuropsychology and Bayesian inference. *Front. Hum. Neurosci.* 12:61. doi: 10.3389/fnhum.2018.00061

Perlovsky, L., and Schoeller, F. (2019). Unconscious emotions of human learning. *Phys. Life Rev.* 31, 257–262. doi: 10.1016/j.plrev.2019.10.007

Pessoa, L. (2010). Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?". *Neuropsychologia* 48, 3416–3429. doi: 10.1016/j.neuropsychologia.2010.06.038

Peters, A., McEwen, B. S., and Friston, K. (2017). Uncertainty and stress: why it causes diseases and how it is mastered by the brain. *Prog. Neurobiol.* 156, 164–188. doi: 10.1016/j.pneurobio.2017.05.004

Pio-Lopez, L., Nizard, A., Friston, K., and Pezzulo, G. (2016). Active inference and robot control: a case study. *J. R. Soc. Interface* 13:20160616. doi: 10.1098/rsif.2016.0616

Pisula, W., and Siegel, J. (2005). Exploratory behavior as a function of environmental novelty and complexity in male and female rats. *Psychol. Rep.* 97, 631–638. doi: 10.2466/pr0.97.2.631-638

Ramstead, M. J., Wiese, W., Miller, M., and Friston, K. J. (2020). *Deep Neurophenomenology: An Active Inference Account of Some Features of Conscious Experience and of their Disturbance in Major Depressive disorder*. Available online at: http://philsci-archive.pitt.edu/18377/ (accessed 30 April, 2021)

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Pers. Soc. Psychol.* 49:95.

Salge, C., and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Front. Robot. AI* 4:25. doi: 10.3389/frobt.2017.00025

Salomon, R. (2017). The assembly of the self from sensory and motor foundations. *Soc. Cogn.* 35, 87–106. doi: 10.1521/soco.2017.35.2.87

Salomon, R., Fernandez, N. B., van Elk, M., Vachicouras, N., Sabatier, F., Tychinskaya, A., et al. (2016). Changing motor perception by sensorimotor conflicts and body ownership. *Sci. Rep.* 6:25847. doi: 10.1038/srep25847

Salomon, R., Progin, P., Griffa, A., Rognini, G., Do, K. Q., Conus, P., et al. (2020). Sensorimotor induction of auditory misattribution in early psychosis. *Schizophrenia Bull.* 46, 947–954. doi: 10.1093/schbul/sbz136

Schoeller, F. (2015). Knowledge, curiosity, and aesthetic chills. *Front. Psychol.* 6:1546. doi: 10.3389/fpsyg.2015.01546

Schoeller, F. (2017). The satiation of natural curiosity. *Int. J. Signs Semiotic Systems* 5, 200516–232707.

Schoeller, F., and Perlovsky, L. (2016). Aesthetic chills: knowledge-acquisition, meaning-making and aesthetic emotions. *Front. Psychol.* 7:1093. doi: 10.3389/fpsyg.2016.01093

Schoeller, F., Perlovsky, L., and Arseniev, D. (2018). Physics of mind: experimental confirmations of theoretical predictions. *Phys. Life Rev.* 25, 45–68. doi: 10.1016/j.plrev.2017.11.021

Schoeller, F., Eskinazi, M., and Garreau, D. (2017). Dynamics of the knowledge instinct: effects of incoherence on the cognitive system. *Cogn. Systems Res.* 47, 85–91. doi: 10.1016/j.cogsys.2017.07.005

Seth, A. (2014). "The cybernetic brain: from interoceptive inference to sensorimotor contingencies," in *MINDS project*, eds T. Metzinger and J. M. Windt (Glastonbury, CT: MINDS).

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K. (2015). "Inference to the best prediction," in *Open MIND*, ed. J. M. Windt (Glastonbury, CT: MIND Group).

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20160007. doi: 10.1098/rstb.2016.0007

Sheridan, T. B. (1988). "Trustworthiness of command and control systems," in *Proceedings of IFAC Man-Machine Systems*, (Oulu), 427–431.

Sheridan, T. B. (2019a). Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Hum. Factors J. Hum. Factors Ergonom. Soc.* 61, 1162–1170.

Sheridan, T. B. (2019b). Individual differences in attributes of trust in automation: measurement and application to system design. *Front. Psychol.* 10:1117. doi: 10.3389/fpsyg.2019.01117

Sheridan, T. B., and Parasuraman, R. (2005). Human-Automation interaction. *Rev. Hum. Factors Ergonom.* 1, 89–129.

Smith, R., Parr, T., and Friston, K. J. (2019). Simulating emotions: an active inference model of emotional state inference and emotion concept learning. *Front. Psychol.* 10:2844. doi: 10.3389/fpsyg.2019.02844

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., et al. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front. Hum. Neurosci.* 10:550. doi: 10.3389/fnhum.2016.00550

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious Cogn.* 17, 219–239.

Van de Cruys, S. (2017). *Affective Value in the Predictive Mind*. Frankfurt: MIND Group.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2019). Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43:e90. doi: 10.1017/S0140525X19001213

Vuorre, M., and Metcalfe, J. (2016). The relation between the sense of agency and the experience of flow. *Conscious Cogn.* 43, 133–142.

Wilkinson, S., Deane, G., Nave, K., and Clark, A. (2019). "Getting warmer: predictive processing and the nature of emotion," in *The value of Emotions for Knowledge* ed. L. Candiotto (Cham: Palgrave Macmillan), 101–119.

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* 12, 957–968. doi: 10.3758/bf03206432

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269:1880.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AO declared a past collaboration with one of the authors FS to the handling Editor.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 21 October 2021
doi: 10.3389/fnsys.2021.675127

Cognitive Neuroscience Meets the Community of Knowledge

*Steven A. Sloman[1], Richard Patterson[2] and Aron K. Barbey[3,4,5,6]**

[1] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, United States, [2] Department of Philosophy, Emory University, Atlanta, GA, United States, [3] Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [4] Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [5] Neuroscience Program, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [6] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Champaign, IL, United States

Cognitive neuroscience seeks to discover the biological foundations of the human mind. One goal is to explain how mental operations are generated by the information processing architecture of the human brain. Our aim is to assess whether this is a well-defined objective. Our contention will be that it is not because the information processing of any given individual is not contained entirely within that individual's brain. Rather, it typically includes components situated in the heads of others, in addition to being distributed across parts of the individual's body and physical environment. Our focus here will be on cognition distributed across individuals, or on what we call the "community of knowledge," the challenge that poses for reduction of cognition to neurobiology and the contribution of cognitive neuroscience to the study of communal processes.

Keywords: community of knowledge, cognitive neuroscience, thinking, collective cognition, social neuroscience

OPEN ACCESS

Edited by:
Rosalyn J. Moran,
King's College London,
United Kingdom

Reviewed by:
Robert Turner,
Max Planck Institute for Human Cognitive and Brain Sciences, Germany
Brandon Turner,
The Ohio State University, United States

***Correspondence:**
Aron K. Barbey
barbey@illinois.edu

Received: *02 March 2021*
Accepted: *30 September 2021*
Published: *21 October 2021*

Citation:
Sloman SA, Patterson R and Barbey AK (2021) Cognitive Neuroscience Meets the Community of Knowledge. Front. Syst. Neurosci. 15:675127. doi: 10.3389/fnsys.2021.675127

THE INDIVIDUAL BRAIN AND COLLECTIVE MIND

A central aim of cognitive neuroscience is to explain how people think, elucidating the representations and processes that allow humans to judge, reason, remember, and decide (Barbey et al., 2021). To achieve this goal, cognitive neuroscientific theories have as a rule made certain fundamental assumptions:

(a) Knowledge is represented in the brain.
(b) Knowledge is represented by the individual.
(c) Knowledge is transferred between individuals.

where "knowledge" is understood broadly—as it usually is in behavioral science—as people's attempts to represent their world, including both observable and latent objects and processes, in ways that support memory, understanding, reasoning, and decision making. It includes beliefs that are more or less justified, and that might correspond to factual truth or not. Evidence to suggest that knowledge is represented in the brain [assumption (a)] may reflect: (1) correlations with neural activity (e.g., spike trains generated by neurons in V1 correlate with the presence and location of edges in the visual environment), (2) causal effects of knowledge on the operation of neural systems (e.g., spike trains generated by neurons in V1 are used by downstream areas for further processing), and/or (3) neural computations applied to manipulate and process knowledge.

Although assumption (a) is typical of theories in the psychological and brain sciences (for reviews, see Gazzaniga et al., 2019; Barbey et al., 2021), it is not universal. Proponents of embodied cognition see knowledge as distributed across the brain, the body, and artifacts used to process information (e.g., Barsalou, 2008) and proponents of cultural psychology sometimes see knowledge as embedded in cultural practices (Duque et al., 2010; Holmes, 2020). But assumptions (b) and (c) are widely shared by disciplines that focus on cognition (for a review, see Boone and Piccinini, 2016). The idea is that what really counts as cognition is mediated by individual processes of reasoning and decision making; that cognitive processing is distinct from interactions with books, the internet, other people, and so on. Moreover, other people are obviously sources of information, but their value for an individual is in the information they transfer. The goal of this manuscript is to question the generality of these assumptions, spell out some of the resulting limitations of the cognitive neuroscience approach, and try to suggest some more constructive directions for the field. Our contention will be that the information processing of any given individual is not contained entirely within that individual's brain (or even their bodies or physical environments). Rather, it typically includes components situated in the heads of others, and that the transfer of information is more the exception than the rule.

Assumption (a) as usually understood implies (b). If knowledge is represented in the brain, then it is represented by individuals. Thus standard neuroimaging methods assess brain activity and task performance within the individual (for a review of fMRI methods, see Bandettini, 2012). According to this view, the neural foundations of the human mind can be discovered by studying the individual brain and identifying common patterns of brain activity across individuals. Thus, by averaging data from multiple subjects, cognitive neuroscience seeks to derive general principles of brain function and thereby reveal the mechanisms that drive human cognition. This approach lies at the heart of modern research in cognitive neuroscience, reflecting a disciplinary aim to generalize beyond the individual to characterize fundamental properties of the human mind using widely held methodological conventions, such as averaging data from multiple subjects, to infer general principles of brain function (Gazzaniga et al., 2019).

Although assumption (a) implies (b), the converse does not also hold. If knowledge is represented by the individual, it need not be represented exclusively within the brain. More importantly, as we will argue, an individual's knowledge not only arises in large part from communal interactions, but also depends on cognitive states of other members of the community. This places limits on the utility of studying individual brains to infer general principles of the collective mind. Our conclusion is decidedly not that cognitive neuroscience makes no contribution to the study of cognition. It is that cognitive neuroscience does not provide a sufficient basis to model cognition. Social neuroscience is an emerging field that addresses part of the problem, as it takes as a central tenet that "brains are not solitary information processing devices" (Cacioppo and Decety, 2011). Nevertheless, the discussions we are aware of within the field of cognitive neuroscience still abide by assumptions (b) and (c).

THE COMMUNITY OF KNOWLEDGE AND THE LIMITS OF THE INDIVIDUAL

We start with assumption (b). Years of research in psychology, cognitive science, philosophy, and anthropology have shown that human cognition is a collective enterprise and is therefore not to be found within a single individual. Human cognition is an emergent property that reflects communal knowledge and representations that are distributed within a community (Hutchins, 1995; Clark and Chalmers, 1998; Wilson and Keil, 1998; Henrich, 2015; Mercier and Sperber, 2017; Sloman and Fernbach, 2017). By "emergent" property we mean nothing elusive or mysterious, but simply certain well-documented properties of groups that would not exist in the absence of relevant properties of individuals, but are not properties of any individual member of the group, or any aggregation of properties of some or all members of the group.

Accumulating evidence indicates that memory, reasoning, decision-making, and other higher-level functions take place across people. The evidence that mental processing is engaged by a community of knowledge is multifaceted (for a review, see Rabb et al., 2019). The claim that the mind is a social entity is an extension of the extended mind hypothesis (Clark and Chalmers, 1998): Cognition extends into the physical world and the brains of others. The point is not that other people know things that I do not; the point is that my knowledge often *depends* on what others know even in the absence of any knowledge transfer from them to me. I might say, "I know how to get to Montreal," when what I really mean is that I know how to get to the airport and the team piloting the aircraft can get from the airport to Montreal. Similarly, one might say that "what makes a car go" is the motor: that's why it's called a "motor," after all. But while a full account will include the engine as a key contributor, the propulsion system is distributed over the engine, drive shaft, the human who turns the key, fuel, a roadway, and more. Changing the boundaries of what has traditionally been considered cognitive processing in an analogous way – from individual brains to interacting communities – perhaps raises questions of who should get credit and who should take responsibility for the effects of an individual's action, but it is nevertheless an accurate description of the mechanisms humans use to process information. Furthermore, as the boundaries for what counts as cognitive processing shift, the operational target for studying the human mind moves beyond the scope of methods that examine performance through the lens of the individual.

Philosophers analyzing natural language illustrate how cognitive processes are extended into the world. The classic analysis is by Putnam (1975), who points out that we often use words whose reference (or denotation or extension) and therefore, according to Putnam, their meaning, is determined by factors outside one's brain or mind (i.e., externalism). One could see Humpty Dumpty as an extreme and defiant internalist:

"When I use a word, it means precisely what I want it to mean, no more and no less" (Carroll, 1872). Putnam's argument is the subject of vigorous and sophisticated but not entirely conclusive debate (Goldberg, 2016; see also Burge, 1979). Nonetheless it is now widely agreed that some form of externalism is at least a necessary part of an explanation of how our everyday terms have their referents (or denotations) and meanings.

The philosopher whom one might call the Godfather of Externalism, Wittgenstein (1973), preferred to draw attention to what he saw as linguistic facts that had been overlooked, above all, that the meaning of words depends on (or is even identical to) their use. Although that bald statement is highly controversial, what matters from our point of view is that the meaning of a word and its correct use depend on collective knowledge that extends beyond the individual, reflecting a social context (Boroditsky and Gaby, 2010). Thus, for a community of knowledge to support meaning and communication, there must be sufficient stability of common usage even as usage typically changes over time. The same holds for sentence meanings, as in, "Zirconium comes after Yttrium in the Periodic Table." The speaker may have long ago forgotten—or never even knew—what exactly Zirconium is and why one thing comes after another in the Periodic Table. Nonetheless the statement has a meaning that has been fixed by the appropriate members of the scientific community, and propagated more-or-less successfully to generations of students. The speaker's statement is true and has that communally established meaning, no matter how confused the speaker may be. Some might distinguish the speaker's meaning from the correct, communally-ordained meaning. That is important in some contexts (e.g., in teaching and in evaluating students), but the point here is that the sentence has a precise meaning established by chemical science, even if that is not precisely what is in the speaker's head, but only in the heads of others.

The same holds of theories. The statement "According to modern chemistry there are more than a hundred elements" is true regardless of how well or poorly the speaker might understand modern chemistry. It is true because "modern chemistry" means the chemical theories agreed upon by socially recognized experts. This holds even if the relevant theories are no longer in the speaker's head, and even if the speaker never understood the theories.

These remarks on social meaning converge with recent work in the emerging discipline of "social epistemology" (Goldman, 1999), the study of knowledge as a social entity. We will speak of "knowledge" in an everyday sense, without entering into the labyrinthine and ultimately inconclusive attempts at definition offered by philosophers from the time of Plato, including what "really constitutes" social knowledge. What matters here is that research within social epistemology demonstrates that successful transmission of knowledge clearly does occur and depends on three general conditions (Goldberg, 2016): (i) social norms of assertion; (ii) reliable means of comprehending what is said (which depend on social norms of meaning and usage); and (iii) a reliable way of telling a reliable source of knowledge from an unreliable one. For reasons we elaborate below, we believe that the role of society in epistemology is not only to transmit knowledge from one individual to another, but to retain knowledge even when it is not transmitted.

Sloman and Fernbach (2017) extended the externalist project well beyond a concern with the meanings of words, to large swathes of conceptual knowledge. Outside their narrow areas of expertise, individuals are relatively ignorant (Zaller, 1992; Dunning, 2011). In any given domain, they know much less than there is to know, but nonetheless do know certain things that others understand more fully. The extent to which we rely on others in this way is often obscured by the fact that people tend to overestimate how much they know about how things work (Rozenblit and Keil, 2002; Lawson, 2006; Fernbach et al., 2013; Vitriol and Marsh, 2018). They overestimate their ability to reason causally (Sloman and Fernbach, 2017). They also overestimate what they know about concept meanings (Kominsky and Keil, 2014) and their ability to justify an argument (Fisher and Keil, 2014) and claim to have knowledge of events and concepts that are fabricated (Paulhus et al., 2003).

The best explanation for our tendency to overestimate how much we know is that we confuse what others know for what we know (Wilson and Keil, 1998). Others know how things work, and we sometimes fail to distinguish their knowledge from our own. The idea is the converse of the curse of knowledge (Nickerson, 1999). In that case, people tend to believe that others know what they themselves know (this is part of what makes teaching hard). In both cases, people are failing to note the boundary among individuals. Circumstances can produce a rude awakening if things go wrong and we suddenly need to understand how to fix them, or if we are otherwise challenged to produce a full explanation either in a real world situation or by a psychologist.

Nonetheless, as Goldman (1999) observes, even a shallow understanding of a concept, idea, or statement can give us valuable practical information. Fortunately, we can know and make use of a good many truths without ourselves possessing the wherewithal to prove them, so long as our limited understanding is properly anchored elsewhere. We develop multiple examples below. Meanwhile, from a very broad perspective, we note that the conceptual web is tangled and immense, containing far more than a mere mortal could store and make sense of Sloman and Fernbach (2017). Thus we are by nature creatures that rely heavily on others to have full understandings of word meanings ("semantic deference" in the philosophical literature) and a more full and secure grasp of ideas, statements, or theories than our own incomplete grasp reflected in our shallow understanding. This dovetails not only with experimental results (Rozenblit and Keil, 2002; Fernbach et al., 2013; Kominsky and Keil, 2014; Sloman and Rabb, 2016), but also with recent anthropological work on culture-gene coevolution showing that cultural accumulation exerted selective pressure for genetic evolution of our abilities to identify and access reliable sources of information and expertise (e.g., Richerson et al., 2010; Henrich, 2015).

At a social level, the fact that knowledge is communal also has a political dimension. As societies develop, group policy and decision-making will depend on the aggregation, coordination, and codification of various sorts of knowledge distributed

across many individuals (e.g., experts in the production, storage, distribution, and preparation of food). There is lively debate among political theorists about whether command and control societies, democracies, or something else can best fulfill the needs and aspirations of its members (Anderson, 2006; Ober, 2008). Is decision-making best served by cloistered experts or through information gathered from non-experts as well? Non-experts presumably have greater access to details of local situations, but attempts to utilize widely distributed knowledge poses greater problems of aggregation and coordination. As Hayek (1945) remarked, the aggregation and deployment of widely distributed information is a central issue for theories of government. However, our interest here is not the relative merits of different forms of government. We mention these issues only to illustrate the far-reaching and pervasive importance of information processing in social networks and by implication the need for a political level of explanation in the understanding of a community of knowledge.

SOCIAL KNOWLEDGE WITHOUT SOCIAL TRANSMISSION: OUTSOURCING

Work on collective cognition points to several ways that individual cognition depends on others (Hemmatian and Sloman, 2018). One is collaboration: Problem-solving, decision-making, memory, and other cognitive processes involve the joint activity of more than one person, and in many contexts mutual awareness of a joint intention to perform some task. Work on collaboration has focused on team dynamics (Pentland, 2012) and group intelligence (Woolley et al., 2010). A second form of cognitive dependence on others, and the one that grounds our argument, is outsourcing: The knowledge people use often sits (or sat) in the head of someone else, someone not necessarily present (or even alive). Frequently, outsourcing requires that we have *access* to outsourced knowledge when the need arises. But often merely knowing we have access is sufficient for practical purposes (e.g., we go to Tahiti assuming we'll find what we need to enjoy ourselves when we're there). On occasion we do access the information, and this requires some type of social transmission. Such transmission comes in the form of social learning of a skill, practice, norm, or theory on the one hand, or in the form of more episodic or *ad hoc* accessing of information for limited, perhaps one-time, use (Barsalou, 1983). A prime example of the former would be an apprentice learning a trade from a master; of the latter, "googling" to find out who won the 1912 World Series. The transmission of information around a social network is a key determinant of human behavior (Christakis and Fowler, 2009).

A key requirement in using information that is sitting in someone else's head is the possession of what we will call epistemic pointers ("epistemic" meaning having to do with knowledge): the conscious or implicit awareness of where some needed information can be found. Sometimes we can envision many potential pathways to an information source, whether direct or indirect, and sometimes very few. Thus we may envision many potential information sources for how to get to Rome (travel agents, friends who have been there), and various pathways by which we might access a given source (e.g., find the phone number of a friend who said she had a good travel agent) but fewer pathways to find out how to get to the rock shaped like an elephant that someone mentioned in passing. Our representations of pointers, to a source or to a step on a pathway to a source, can be partial and vague, providing little or no practical guidance ("some physics Professor knows it"), or full and precise ("it's in Einstein's manuscript on the special theory of relativity"). If we are completely clueless, we can be said to lack pointers and pathways, and simply have a placeholder for information. The evidence of human ignorance that we review below leads us to suspect that the vast majority of the knowledge that we have access to and use is in the form of placeholders.

SETTING THE STAGE: COLLABORATION

The centrality of collaboration for human activity derives from the fact that humans are unique in the cognitive tools they have for collaboration. Tomasello and Carpenter (2007) make the case that no other animal can share intentionality in the way that humans can in the sense of establishing common ground to jointly pursue a common goal, and a large body of work describes the unique tools humans have to model the thoughts and feelings, including intentions and motivations, of those around them (e.g., Baron-Cohen, 1991).

The role of collaboration in specifically cognitive performance has been most fully studied in memory. Wegner et al. (1991) report some of the early work showing that groups, especially married couples, distribute storage demands according to relative expertise. They call these "transactive memory systems." Theiner (2013) argues that transactive memory systems reflect emergent group-level memories, providing evidence that: (i) members of a transactive memory system are not interchangeable (because each member makes unique contributions to the group); (ii) if members are removed from the group, the system will no longer function (omitting essential components of the group-level memory); (iii) the disassembly and reassembly of the group may impair its function (for example, when members of the group no longer understand the distribution of knowledge within the system and what information they are responsible for knowing); and (iv) cooperative and inhibitory actions among members are critical (given the interactive and emergent nature of transactive memories) (for a review, see Meade et al., 2018). Wilson (2005) claims that these properties of a transactive memory system have important political consequences as they affect the commemoration and memorialization of politically relevant events and culturally important origin stories that shape nationalism and attitudes toward human rights and other issues. Memory systems play a critical role in communities.

Further evidence for the importance of collaboration in thought comes from naturalistic studies of group behavior. The seminal work was conducted by Hutchins (1995). He offered a classic description of navigating a Navy ship to harbor, a complex and risky task. The process involves multiple people

contributing to a dynamic representation of the ship's changing location with reference to a target channel while looking out for changing currents and other vessels. Various forms of representation are used, all feeding into performance of a distributed task with a common goal. Sometimes the common goal is known only by leadership (in the case of a secret mission, say). Nevertheless, successful collaboration involves individuals pursuing their goals so as to contribute to the common goal. Many of the tasks we perform everyday have this collaborative nature, from shopping to crossing the street. If a car is coming as we cross, we trust that the driver won't accelerate into us, and the more assertive street crossers among us expect them to slow down in order to obtain the common goal of traffic flow without harm to anyone. Banks and Millward (2000) discuss the nature of distributed representation and review data showing that distributing the components of a task across a group so that each member is a resident expert can lead to better performance than giving everyone the same shared information. Hutchin's nautical example illustrates this, insofar as some essential jobs require multiple types of expertise. Other jobs might not require this, so that crew members may substitute for one another, because all of them have the same basic information or skill level needed for the job. Often in real life there will be a mix, so that the task occupies an intermediate position relative to Banks and Millward's two types of group (i.e., diverse local experts versus all group members having the same knowledge). Work on collective intelligence also provides a good example of emergent group properties, illustrating how collective problem-solving relies more on collaboration and social interconnectedness than on having individual experts on the team (Woolley et al., 2010).

COLLABORATION AND NEUROSCIENCE: THE CASE OF NEURAL COUPLING

Research in cognitive neuroscience has not ignored these trends in the study of cognition. An emerging area of research investigates the communal nature of brain networks, examining how the coupling of brain-to-brain networks enables pairs of individuals or larger groups to interact (Montague et al., 2002; Schilbach et al., 2013; Hasson and Frith, 2016). These studies deploy a generalization of neuroimaging methods, applying techniques that were once used to assess intra-brain connectivity (i.e., within the individual) to examine inter-subject connectivity (i.e., between different subjects; Simony et al., 2016). This can be achieved through experiments in which brain activity within multiple participants is simultaneously examined (i.e., "hyperscanning;" Montague et al., 2002) or analyzed *post hoc* (Babiloni and Astolfi, 2014). Such approaches have been applied to assess brain-to-brain communication dynamics underlying natural language (e.g., Schmalzle et al., 2015). Recently, researchers have placed two people face-to-face in a single scanner to examine, for example, the neural mechanisms underlying social interaction (e.g., when people make eye contact; for a review, see Servick, 2020). The situation – very noisy and now also very crowded – does not score high on ecological validity. Also, it is hard to see how one could scale this approach up to study larger groups (big scanners, little participants?). Nonetheless this is a reasonable place to start, and here, as with hyperscanning and retrospective analysis of neuroimaging data, one might well secure suggestive results. So although the examination of brain-to-brain networks is rare in cognitive neuroscience, with only a handful of studies conducted to date (for a review, see Hasson and Frith, 2016), this approach represents a promising framework for extending cognitive neuroscience beyond the study of individuals to an investigation of dyads, groups, and perhaps one day to larger communities.

This approach has set the stage for research on the neural foundations of communal knowledge, investigating how cognitive and neural representations are distributed within the community and how information propagates through social networks, for example, based on their composition, structure, and dynamics (for a review, see Falk and Bassett, 2017; for a discussion of hyperscanning methods, see Novembre and Iannetti, 2020; Moreau and Dumas, 2021). Evidence from this literature indicates that the strength of the coupling between the neural representation of communication partners is associated with communication success (i.e., successful comprehension of the transmitted signal; Stephens et al., 2010; Silbert et al., 2014; Hasson and Frith, 2016). For example, the degree of brain-to-brain synchrony within networks associated with learning and memory (e.g., the default mode network) predicts successful comprehension and memory of a story told among communication partners (Stephens et al., 2010). Indeed, evidence indicates that people who are closely related within their social network (i.e., individuals with a social distance of one) demonstrate more similar brain responses to a variety of stimuli (e.g., movie clips) relative to individuals who share only distant relations (Parkinson et al., 2017). Research further suggests that the efficiency of inter-subject brain connectivity increases with the level of interaction between subjects, providing evidence that strong social ties predict the efficiency of brain-to-brain network coupling (Toppi et al., 2015; for a discussion of the timescale of social dynamics, see Flack, 2012).

THE MAIN EVENT: OUTSOURCING

A community of knowledge involves more than coupling. We do collaborate, and we engage in joint actions involving shared attention, but we also make use of others without coupling: We outsource to knowledge housed in our culture, beyond the small groups we collaborate with. In the best cases, we outsource to experts. A great many people know that the earth revolves around the sun, but only a much smaller number know how to show that. Both sorts of people are part of a typical community of knowledge, and both are, by community standards, said to know that the earth revolves around the sun. This holds even though the non-expert does not know who the experts are, does not remember how she came to have that knowledge, and does

not know what observations and reasoning show that our solar system is heliocentric.

Outsourcing in some circumstances can make us vulnerable to a lack of valuable knowledge. Henrich (2015) describes how an epidemic that killed off many older and more knowledgeable members of the Polar Inuit tribe resulted in the tribe losing access to much of its technology: Weapons, architectural features of their snow homes, and transportation (e.g., a particular type of kayak). Knowledge about how to build and use these tools resided in the heads of those lost members. Without them, the remaining members of the tribe were unable to figure out how to build such tools, and were forced to resort to less effective means of hunting, staying warm, and traveling. The issue here is not collaboration. Tool users were not cognitively coupling with the tool providers. Rather, they were accessing and making use of the latter's knowledge without acquiring it, in this case outsourcing both the expertise and the production of vital artifacts. Assumptions that individuals had been able to rely on (i.e., that they would have access to a tool for obtaining food) no longer held. The problem was that the younger members of the tribe had outsourced their knowledge to others who were no longer available. Anthropologists have documented numerous cases of loss of technology through death of the possessors of a society's specialized knowledge, or through isolation from formerly available knowledge sources (e.g., Henrich and Henrich, 2007). By the same token, a community can add new expertise by admitting (or forcibly adding) new members with special skills (e.g., Weatherford, 2005).

Sometimes we are aware that we are outsourcing, for instance when we explicitly decide to let someone else do our cognitive work for us (as one lets an accountant file one's taxes). In such cases, we explicitly build a pointer, a mental representation that indicates the repository of knowledge we do not ourselves fully possess and that anchors the shallow or incomplete knowledge we do possess. We have a pointer to an accountant or tax lawyer (whether to a specific person or just to a "tax preparer to be determined"), just in case we are audited.

But often we outsource without full awareness, acting as if we have filled gaps in our knowledge even though no information has been transferred. Our use of words is often licensed by knowledge only others have, our explanations often appeal to causal models that sit in the heads of scientists and engineers, and our political beliefs and values are inherited from our spiritual and political communities. More generally, people's sense of understanding, reasoning, decision-making, and use of words and concepts are often outsourced to others, and often we do not know whom we are outsourcing to, or even that we are doing it. For instance, when we say "*they* landed on the moon," most of us have little idea who *they* refers to, and often lack conscious awareness that we don't know who *they* were. Or we say, "We know that Pluto is not strictly speaking a planet." We know that much on reliable grounds. What little we know is anchored by the possibility of transmission (direct or perhaps very indirect) from communal experts; specifically, the scientists who set the criteria for planethood, and who know whether Pluto qualifies and on the basis of what evidence. Again, it is highly advantageous to be able to outsource – and in fact necessary – since we can't all master full knowledge of all the crafts, skills, theoretical knowledge, and up-to-date-details of local situations that we need or might need to navigate our environment.

Moreover, people believe they understand the basics of helicopters, toilets, and ballpoint pens even when they do not (Rozenblit and Keil, 2002). Fortunately, others do. In addition, the knowledge that others do increases our sense of understanding not only of artifacts, but of scientific phenomena and political policies (Sloman and Rabb, 2016; Rabb et al., 2019). In fact, just having access to the Internet also increases our sense of understanding even when we are unable to use it (Fisher et al., 2015). These findings cannot be attributed to memory failures because, in the vast majority of cases, the relevant mechanisms were never understood. And the studies include control conditions to rule out alternative explanations based on self-presentation effects and task demands. What they show is that mere access to information increases our sense of understanding. This suggests our sense of understanding reflects our roles as members of a community of knowledge, and suggests that we maintain pointers to or placeholders for information that others retain. The fact that access causes us to attribute greater understanding to ourselves implies that our sense of understanding is inflated. This in turn implies that we fail to distinguish those pointers or placeholders from actual possession of information; we don't know that we do not really know how artifacts like toilets work, but the awareness that others do leads us to think we ourselves do, at least until we are challenged or we land in a situation demanding genuine expertise (Call the plumber now!).

More evidence for this kind of implicit outsourcing comes from work on what makes an explanation satisfying. People find explanations of value even if they provide no information, as long as the explanations use words that are entrenched in a community. For example, Hemmatian and Sloman (2018) gave subjects a label for a phenomenon (e.g., "Carimaeric") and told them that the label referred to instances with a specific defining feature (e.g., stars whose size and brightness varied over time). Then the label was used as an explanation for the defining property (someone asked why a particular star's size and brightness varied over time and was told that it's because the star is Carimaeric). Subjects were asked to what extent the explanation answered the question. They answered more positively if the label was entrenched within a community than if it was not. Similar findings have been obtained using mental health terms, even among mental health professionals (Hemmatian et al., 2019). In these cases, there is no coupling between the unidentified community members who use the explanation and the agent. There is merely the heuristic that the fact that others know increases my sense of understanding. This heuristic is so powerful that it operates even when others' knowledge has no informational content.

Some of the clearest evidence for this heuristic comes from the political domain. We often take strong stances on issues that we are ignorant about. These authors believe strongly in anthropogenic climate change despite being relatively ignorant of both the full range of evidence and the mechanism for it. We rely on those scientists who study such things. Political issues tend

to be complex and we need to rely on others, at least in part, to form and justify our opinions. In a representative democracy, for instance, we try to be informed on key issues, but rely on specialized committees to investigate matters more thoroughly. For better or for worse, individual support for policies, positions, and leaders comes largely from partisan cues rather than non-partisan weighing of evidence (Cohen, 2003; Hawkins and Nosek, 2012; Anduiza et al., 2013; Han and Federico, 2017; Van Boven et al., 2018). A growing body of evidence indicates that partisan cues determine how we understand events (Jacobson, 2010; Frenda et al., 2013; but see Bullock et al., 2015) and even whether we take steps to protect ourselves from infectious disease (Geana et al., 2021)[1]. Marks et al. (2019) show that people use partisan cues to decide whose advice to follow in a competitive game even when they have objective evidence about who the better players are. When evaluating data, we are often more concerned with being perceived as good community citizens by acceding to our community's mores than we are with making accurate judgments (Kahan et al., 2011). Such a bias has a rationale if it maintains community membership, and membership is deemed more important than being correct.

Outsourcing knowledge, including the choice of whom to outsource to, is a risky affair. One must estimate what the source does and does not know, their ability to transmit information, and whether their interests align with yours. One must determine how much to trust potential sources of information. Outsourcing, whether influenced by partisan bias or not, is a direct consequence of the human need and tendency to construct pointers to knowledge that other people store.

The basic features of how a community holds knowledge—relative ignorance associated with epistemic pointers to expertise—apply to both social information and disinformation, to well-grounded knowledge, as well as fervently held nonsense perpetrated by unreliable sources. Community norms about what counts as knowledge, and as a reliable pathway of knowledge transmission, may vary greatly: One subculture will require, for some subject matters, scientific expertise on the part of an ultimate source, along with reliable paths of transmission of scientific knowledge, paths often institutionalized, as with schools or trade unions and their certifications. Another subculture will consider God the ultimate source of understanding in important areas, and divine revelation, or the word of officially ordained spokespersons, as appropriate paths of dissemination.

Thus the role of our social networks goes beyond actively sharing information. We use them to represent and process information, such that the network itself serves as an external processor and storage site. We trust others to maintain accurate statistics, to distil news, to total our grocery bill, help us fill out our tax forms, and to tell us what position to take on complex policy. In all such tasks, representation and processing of essential information does not in general occur in individual brains. They do not occur in individual brains even if we allow that those brains are coupled within a social network. Representation and processing occur over a larger portion of an encompassing network, and potentially over the entire network, branching out to include our sources, our sources' sources, and any intermediaries such as books, the internet, or other people, along the paths of transmission.

OUTSOURCING IN COGNITIVE NEUROSCIENCE: CONSTRUCTING EPISTEMIC POINTERS

To explain phenomena associated with outsourcing, we cannot appeal to coupling, because coupling requires specification of who is coupling with whom. To explain outsourcing, cognitive neuroscientists must appeal to a different theoretical construct: Neural pointers or placeholders, representations in the brain that act as pointers to knowledge held elsewhere. The work in cognitive neuroscience that most directly addresses the mechanisms of outsourcing concerns how the representation of knowledge relates to affiliation, on whom we trust to retain reliable knowledge. Putting aside the role of trust in institutions, social neuroscience research examining trust in more personal contexts indicates that trust and cooperation are mediated by a network of brain regions that support core social skills, such as the capacity to infer and reason about the mental states of others (for reviews, see Adolphs, 2009; Rilling and Sanfey, 2011). This work provides the basis for future research investigating how the neurobiology of trust contributes to the representation and use of outsourcing in collective cognition. To do so, however, the field will need to move beyond the use of "isolation paradigms" in which subjects observe others whom they might or might not then trust (Becchio et al., 2010). In such cases, subjects neither participate in direct social interaction with potential objects of trust nor outsource their own reasoning to others (Schilbach et al., 2013). Such observation is seldom the sole basis of epistemic pointers, and often is not involved at all. Instead, pointers typically depend on cues that reflect how third parties or the community as a whole regard a potential source. This can involve informal gossip or more institutionalized "rating systems" and reviews, where the latter will bring us back to social institutions. So there is a vast arena, virtually unexplored by social neuroscience, starting with the origin and nature of the neural mechanisms that serve as pointers to communal knowledge.

THE IRREDUCIBILITY OF THE COMMUNITY OF KNOWLEDGE

The implication of our discussion is that many activities that seem solitary—like writing a scientific paper—require a cultural community as well as the physical world now including the Internet (to ground language, to support claims, to provide inspiration and an audience, etc.). Does this mean there is no solely neurobiological representation for performing such

[1]Geana, M., Rabb, N., and Sloman, S. A. (2021). *Walking the Party Line: The Growing Role of Political Ideology in Shaping Health Behavior in the United States.* Manuscript under review.

tasks? Perhaps neurobiological reduction can be accomplished by giving up on the idea of reduction to a single brain, and instead appeal to reduction to a network of brains (Falk and Bassett, 2017). Perhaps a broader view of cognitive neuroscience as the study of information processing in a social network of neural networks can overcome the challenge posed for cognitive neuroscience by the community of knowledge. Can networks of individuals processing together be reduced to networks of brains interconnected by some common resource, perhaps some form of neural synchrony?

We believe the answer is "no." For one thing, the relevant social network is frequently changing, as is membership in groups addressing different problems (for climate change, it involves climate scientists but for predicting football scores, it involves football fans). So there are no fixed neurobiological media to appeal to. This might seem to be irrelevant, as the goal of cognitive neuroscience is not to reduce cognition to a group of specific brains. Rather, one studies specific brains in order to find general patterns of activity that occur in different brains. But this is precisely the problem; namely, the general pattern may not capture specific properties exhibited by the individual. Generalization from the group to the individual depends on equivalence of the mean and variance at each level; an equivalence that has increasingly been called into question (Fisher et al., 2018). The same problem will almost certainly arise with generalizations about multiple groups' performance of a given task. Indeed the problem may be much worse, as changing group membership may introduce even greater variation across groups of the patterns of interaction that produce a group's performance.

Changes in membership will not just mean changes in the attributes and resources the members bring to the group, but also – and more strikingly – potentially very large differences in the way members interact, even if they happen to produce the same result (e.g., if they forecast the same football score as another group whose members interacted in their own, different way in arriving at that prediction). Studies of group dynamics and organizational behavior recognize that many factors affect the efficiency and result of group collaboration: the relative dominance of discussion by some particular member(s), the timidity of others, the motivations of members, the level of experience and expertise of the members, the level of relevant knowledge about the particular teams involved, the stakes involved in making a good prediction, time limitations, the degree of synergy among team members, size of the group, form of discussion used (Hirst and Manier, 1996; Cuc et al., 2006), demographic makeup of the members, and so on. Different fans, or even the same fans on different occasions, can arrive at the same score forecasts for the same game by an unlimited number of patterns of interaction. This not only produces the problem of multiple realization (of a type of group performance on a given task) on a grand scale, but indicates that there will be no tolerably definite and generalizable pattern of group dynamics that applies to particular groups addressing the same given task. Hence there is no one general pattern, or even manageable number of patterns, to be reduced to neuroscience.

On a more positive note, research in group dynamics and organizational behavior has, as just noted, identified numerous factors that enter into group performance. So cognitive neuroscience (social and individual) can, by drawing on that research, investigate the neural underpinnings of types of factors such as trust, mind-reading capacities, and many others that drive different forms of group interaction, and this will be essential for an account of group cognition if such an account is ever to be had. But that is a far cry from reducing group behavior to any variety of neuroscience.

GROUP INTELLIGENCE AND INVENTIVENESS

Anthropological and psychological research, in the lab and in the field, strongly reinforces the point: group intelligence and group inventiveness are not just the properties of an individual (such as the smartest or most inventive member of the group), or an average of the members' properties, or an aggregate of the members' individual cognitive properties (Woolley et al., 2010). They are sometimes quite surprising properties that emerge from interactions among members of the group, in some cases as a matter of learning, sometimes just from a repeated exchange of ideas, sometimes from a group of initially equal members, sometimes from a group with one or two initial stand outs. The effect of group interaction can be positive or negative depending on the motivations, personal traits, group camaraderie and various situational constraints (e.g., time limitations, availability of paper and pencil, food, and rest).

The moral is that examination of the brains of group members will not reveal or predict precisely how the group as a whole will perform, nor through what complex pattern of interaction or mechanisms it arrived at a given result. Even in a relatively small group there will be an enormous number of interactions that might produce any given result, and that number will increase exponentially with any increase in group size, not to mention the introduction of other potentially influential factors.

Thus there is no way to identify any particular neurobiological pattern (or manageably small number of patterns) across brains as *the* way(s) in which groups produce new knowledge, or even the way the same group functions on different occasions or with regard to different sorts of cognitive tasks. Put another way, even if we could find out through observation, self-report, or fMRI conducted in everyone, that specific members of a given group engaged in certain specific types of interaction with other specific members, and we were able to reduce that to neurobiological terms, we would not be able to say more than that this is one of innumerable ways a particular group result might be realized in a particular social and physical context. An open-ended list of possible realizations at the psychological or behavioral level does not support a reduction of this bit of psychological description to cognitive neuroscience even if it tells us a lot about what goes into that performance. Note once again that we need functional descriptions, which will themselves be complex and predictive of behavior in only a limited way. Functional descriptions will, as with individual psychology and neuroscience, provide essential guidance and support for social neuroscience, and potentially draw on insights from neuroscience.

JUSTIFICATION AND COMMUNAL NORMS

We saw earlier that within a community of knowledge most of what we know is anchored in the heads of people doing scientific, technical, and other sorts of intellectual work, or in the knowhow of expert mechanics, electricians, potters, and so on. Thus, most of an individual's knowledge is just more or less shallow understanding or very limited practical knowhow, along with a more-or-less precise pointer to expert knowledge (Rabb et al., 2019). For instance, we know that "smoking causes lung cancer" but most of us are not sure why. So the neurobiological representations under study are really mostly pointers to knowledge that experts have or to pathways of transmission by which we can reliably access that information. Hence, the network that anchors much of our knowledge about the causal structure of the world is actually a network that sits across brains, not within a brain: It is not an aggregate of brain contents, but a pattern of interactions among brains with certain contents. Because it is the contents that are important, and not the specific brains, there are an unlimited number of patterns of interactions that would generate and maintain the same causal beliefs.

But the actual justification for those beliefs is more systematic than that. We have seen that it depends on community norms for attributing knowledge and associated institutions of knowledge certification. Within a given community, whatever complies with those norms qualifies as knowledge. Some communities may have rather eccentric norms, and regard some things as general knowledge that another community regards as wild-eyed conspiracy theory (issues of fake news and slander come to mind). Accordingly, an account of most of our knowledge will need to include the role of such social institutions and norms. I can legitimately claim to know that the sun does not revolve around the earth, that anthropocentric climate change is real, that the Pythagorean Theorem is true, and a great many other things I "learned in school," even if I cannot myself produce proofs for any of them, or even say precisely what they amount to (Note that this is different from the case in which I could produce a proof if I sat down and tried to work one out). I know these things because they are known by recognized knowledge sources and I got them from socially recognized reliable transmitters of knowledge. This holds even if I can't now remember where I learned it and am not capable of coming up with the evidence or proofs that sit in the heads of others.

My indirect and usually very superficial knowledge is anchored in the social network of experts and paths of transmission. Similarly, even the knowledge of experts is typically anchored in large part in that of other experts, as architects rely on results in materials science, industrial design, designers and manufacturers of drafting tables and instruments, and so on. Again, an enormous amount of anyone's knowledge exists only by way of a larger community of cognizers and their interactions. These aspects of knowledge—including knowledge worked out in the privacy of my study or laboratory—are "knowledge" only by virtue of being anchored in a larger social network, independently of the particular neurobiology they are grounded in.

Consider a team of researchers writing a manuscript together. A complete account of collaboration and outsourcing involved in joint manuscript writing would have to include not only the brains of the authors, but also those whose evidence or testimony provides the support for claims made in the manuscript. If the manuscript presents findings summarizing a report, then the network would have to include the brains of everybody who wrote the report, or perhaps only those who contributed relevant parts. But how would you decide whose brain is relevant? It would depend on whether relevant knowledge was referenced in the manuscript. In other words, the structure of the knowledge is necessary to determine the relevant source and corresponding neural network to represent that knowledge. The knowledge would therefore not be reducible to a neural network, because identifying the network would depend on the knowledge.

Anyone attempting to describe the cross-brain neural network involved in writing a given manuscript, in the relevant processing and transmission (or lack thereof) of various sorts of information from multiple diverse sources, would not know which brains to look at, or what to look for in different brains, without already being able to identify how each bit of information in the manuscript is grounded. But even if we could identify *a posteriori* the network of brains or profiles of brain activity pertinent to a given piece of collaborative writing, we would be no further in explaining how or why the article came to be written. The reason that some ideas enter into a representation is because they elaborate on or integrate the representation in a more or less coherent way. One reason a report gets cited in a manuscript is that it supports or illustrates some informational point. If there is resonance among neural networks, it is because the information they represent is resonant; the neural networks are secondary. The knowledge held by the community is driving; any emergent neural networks are coming along for the ride.

At the beginning of this essay, we stated three widely-held assumptions in cognitive neuroscience that are inconsistent with facts about what and how people know. Our aim is not to diminish the important contributions of cognitive neuroscience. The assumptions we stated do hold for a variety of critical functions: Procedural knowledge is held in individual brains (or at least individual nervous systems in interaction with the world), and people obviously retain some symbolic knowledge in their individual brains. Moreover, common sense is enough to indicate that knowledge at a basic-level (Rosch, 1978) is regularly transferred between individuals. But far more symbolic knowledge than people are aware of is held by others – outside the individual's brain. Thus, the purpose of much of cognitive neuroscience, to reduce knowledge to the neural level, is a pipe dream. The fact of communal knowledge creates a key limitation or boundary conditions for cognitive neuroscience.

SUMMARY AND IMPLICATIONS

We have elaborated a theory of the community of knowledge, identifying as primary components outsourcing and collaboration, along with an hypothesis about how we construct

BOX 1 | Cognitive neuroscience meets the community of knowledge.

Our understanding of how the world works is limited and we often rely on experts for knowledge and advice. One way that we rely on others is by *outsourcing* the cognitive work and task of reasoning to experts in our community. For example, we believe that "smoking causes lung cancer" even though many of us have little understanding of why this is the case. Here, we simply appeal to knowledge and expertise that scientists within our community hold.

And we behave in a manner that is consistent with knowing this information. We believe that smoking would elevate the risk of lung cancer; if a person were diagnosed with lung cancer, we would suppose they were a smoker; and we choose not to smoke because of the perceived cancer risk. But, again, an explanation for why "smoking causes lung cancer" is something that most of us do not know or understand. Our limited understanding simply relies on experts in the community who have this knowledge; we outsource the cognitive task of knowing and rely on experts for advice.

It may appear that this example is a special case and that we rarely outsource our knowledge to others. But, in fact, we do this all the time. Think of how well people understand principles of science, medicine, philosophy, history, and politics, or how modern technology works. We often have very little knowledge ourselves and instead rely on others to understand, think, reason, and decide. This reliance reflects how our individual beliefs are grounded in a *community of knowledge*.

By appealing to the community, we can ground our limited understanding in expert knowledge, scientific conventions, and normative social practices. Thus, the community justifies and gives meaning to our shallow knowledge and beliefs. Without relying on the community, our beliefs would become untethered from the social conventions and scientific evidence that are necessary to support them. It would become unclear, for example, whether "smoking causes lung cancer," bringing into question the truth of our beliefs, the motivation for our actions, and no longer supporting the function that this knowledge serves in guiding our thought and behavior. Thus, to understand the role that knowledge serves in human intelligence, it is necessary to look beyond the individual and to study the community.

In this article, we explore the implications of outsourcing for the field of cognitive neuroscience: To what extent is cognitive neuroscience able to study the communal nature of knowledge? How would standard neuroscience methods, such as fMRI or EEG, capture knowledge that is distributed within the community? In the case of outsourcing, knowledge is not represented by the individual and knowledge is not transferred between individuals (i.e., it is the expert(s) who hold the knowledge). Thus, to study outsourcing, cognitive neuroscience would need to establish methods to identify the source of knowledge (i.e., who has the relevant information within the community?) and characterize the socially distributed nature of brain network function (e.g., what is the neural basis of outsourcing and the capacity to refer to knowledge held in the community?).

In this article, we identify the challenges this poses for cognitive neuroscience. One challenge is that representing the source of expertise for a given belief is not straightforward because expertise is time and context dependent, may rely on multiple members of the community, and may even depend on experts that are no longer alive. Another challenge is that outsourcing may reflect emergent knowledge that is distributed across the community rather than located within a given expert (e.g., knowledge of how to operate a navy ship is distributed across several critical roles; Hutchins, 1995). Standard methods in cognitive neuroscience, such as fMRI or EEG, are unable to directly assess knowledge distributed in the community because they are limited to examining the brains of individuals (or, at most, very small groups).

Thus, we argue that the outsourcing of knowledge to the community cannot be captured by methods in cognitive neuroscience that attempt to localize knowledge within the brain of an individual. We conclude that outsourcing is a central feature of human intelligence that appears to be beyond the reach of cognitive neuroscience.

epistemic pointers to potential sources of knowledge, whether those sources be people to whom we outsource knowledge or with whom we might collaborate. Our hypothesis places limits on the power of cognitive neuroscience to explain mental functioning (**Text Box 1**). Cognitive neuroscience has often focused on tasks that, at least on their face, are performed by individuals (cf., Becchio et al., 2010; Schilbach et al., 2013). But the limited predictive power of these tasks for human behavior may reflect the fact that these tasks and methods do not capture normal human thinking and may explain some of the limited replicability and generalization of fMRI findings (Turner et al., 2019). People devote themselves to tasks that involve artifacts and representational media designed by other people, to issues created by other people, to ideas developed by and with other people, to actions that involve other people, and of course to learning from sources outside themselves. None of these tasks are amenable to a full accounting from cognitive neuroscience.

Furthermore, our appeal to collective knowledge serves to reinforce the multiple realizability problem (Marr, 1982), allowing functional states to operate over complex and dynamic social networks. Whatever neural representations correspond to a bit of knowledge, they are tied to my belief by virtue of a functional relation (a placeholder in my brain that expresses the equivalent of "experts believe this!"), along with the existence of a reliable pedigree for that belief, not simply because my brain is part of a larger neural network. Functional states reflect communal knowledge. Because the human knowledge system is distributed across people, the parts of it that are anchored in others' knowledge are beyond the reach of cognitive neuroscience.

In sum, the community of knowledge hypothesis implies that it's a mistake to think of neurobiology as sitting beneath and potentially explaining the cognition that constitutes the emergent thinking in which groups and communities engage. And that's most thinking. It also implies that components of that socially distributed cognitive system cannot in principle be defined in terms of or eliminated in favor of neurobiology.

Notice that our argument against reductionism has nothing to do with the nature of consciousness, the target of many such arguments (Searle, 2000; Dennett, 2018). In our view, this is a virtue because consciousness has escaped serious scientific analysis and therefore provides little ground for a serious scientific argument. The representations entailed by collective cognition, in contrast, can be analyzed. In principle, the representations involved in (say) designing a complex object may be abstract in the sense that they reflect interactions among knowledge stored in multiple brains, as well as the physical and virtual worlds, but they are describable nonetheless. As such, the emergent features of human cognition that we are advocating are well-documented and well-established as subjects of fruitful scientific research.

Our argument does have positive implications about how to make progress in cognitive neuroscience. To mention only some of the most basic of these, it suggests that our models of information processing for most tasks should focus on communal, not individual, representations. Because most of

what we know and reason about is stored outside our heads, our models should not be exclusively about how we represent content, but also about how we represent pointers toward knowledge that is housed elsewhere. Because our actions are joint with others, models of information processing require not only a notion of intention, but a notion of shared intention (Tomasello et al., 2005). Finally, models of judgment that apply to objects of any complexity need to address how we outsource information, not just how we aggregate beliefs and evidence.

CONCLUSION

The goal of this article is to focus cognitive neuroscientists on important facts about cognitive processing that have been neglected, and that, if attended to, would facilitate the project of cognitive neuroscience. Greater understanding of how people collaborate would help reveal how neural processing makes use of group dynamics and affiliation, and it would support a more realistic model of mental activity that acknowledges individual limitations. Greater understanding of how people outsource would help reveal the actual nature and limits of neural representation, and shed light on how people organize information by revealing how they believe it is distributed in the community and the world. And greater appreciation of the emergent nature of knowledge in society would help us recognize the limits of cognitive neuroscience, that the study of the brain alone cannot reveal the representations responsible for activities that involve the community. Thus, we join the call for a new era in cognitive neuroscience, one that seeks to establish explanatory theories of the human mind that recognize the communal nature of knowledge and the need to assess cognitive and neural representations at the level of the community – broadening the scope of research and theory in cognitive neuroscience by recognizing how much of what we think depends on other people.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract 2014-13121700004 to the University of Illinois at Urbana-Champaign (PI: AB) and the Department of Defense, Defense Advanced Research Projects Activity (DARPA), via Contract 2019-HR00111990067 to the University of Illinois at Urbana-Champaign (PI: AB).

ACKNOWLEDGMENTS

We thank Robert N. McCauley, Nathaniel Rabb, and Adam H. Russell for improving our argument and Kyle Baacke for assistance in the preparation of the manuscript.

REFERENCES

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* 60, 693–716. doi: 10.1146/annurev.psych.60.110707.163514

Anderson, E. (2006). The epistemology of democracy. *Episteme* 3, 8–22. doi: 10.1353/epi.0.0000

Anduiza, E., Gallego, A., and Munoz, J. (2013). Turning a blind eye: experimental evidence of partisan bias in attitudes toward corruption. *Comp. Political Stud.* 46, 1664–1692. doi: 10.1177/0010414013489081

Babiloni, F., and Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: past, present and future. *Neurosci. Biobehav. Rev.* 44, 76–93. doi: 10.1016/j.neubiorev.2012.07.006

Bandettini, P. A. (2012). Twenty years of functional MRI: the science and the stories. *Neuroimage* 62, 575–588. doi: 10.1016/j.neuroimage.2012.04.026

Banks, A. P., and Millward, L. J. (2000). Running shared mental models as a distributed cognitive process. *Br. J. Psychol.* 91(Pt 4), 513–531. doi: 10.1348/000712600161961

Barbey, A. K., Karama, S., and Haier, R. J. (2021). *Cambridge Handbook of Intelligence and Cognitive Neuroscience*. Cambridge: Cambridge University Press.

Baron-Cohen, S. (1991). "Precursors to a theory of mind: understanding attention in others," in *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, Vol. 1, Ed. A. Whiten (Hoboken, NJ: Basil Blackwell), 233–251.

Barsalou, L. W. (1983). Ad hoc categories. *Mem. Cogn.* 11, 211–227. doi: 10.3758/bf03196968

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Becchio, C., Sartori, L., and Castiello, U. (2010). Toward you: the social side of actions. *Curr. Dir. Psychol. Sci.* 19, 183–188. doi: 10.1177/0963721410370131

Boone, W., and Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese* 193, 1509âĂŞ1534. doi: 10.1007/s11229-015-0783-4

Boroditsky, L., and Gaby, A. (2010). Remembrances of times East: absolute spatial representations of time in an Australian aboriginal community. *Psychol. Sci.* 21, 1635–1639. doi: 10.1177/0956797610386621

Bullock, J. G., Gerber, A. S., Hill, S. J., and Huber, G. A. (2015). Partisan bias in factual beliefs about politics. *Q. J. Polit. Sci.* 10, 519–578. doi: 10.1561/100.00014074

Burge, T. (1979). Individualism and the mental. *Midwest Stud. Philos.* 4, 73–121. doi: 10.1111/j.1475-4975.1979.tb00374.x

Cacioppo, J. T., and Decety, J. (2011). *An Introduction to Social Neuroscience. Oxford Handbook of Social Neuroscience*. Oxford: Oxford University Press.

Carroll, L. (1872). *Through the Looking-Glass and What Alice Found There*. Chicago: W.B. Conkey Co.

Christakis, N. A., and Fowler, J. H. (2009). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York, NY: Little, Brown Spark.

Clark, A., and Chalmers, D. (1998). The extended mind (active externalism). *Analysis* 58, 7–19.

Cohen, G. L. (2003). Party over policy: the dominating impact of group influence on political beliefs. *J. Pers. Soc. Psychol.* 85, 808–822. doi: 10.1037/0022-3514.85.5.808

Cuc, A., Ozuru, Y., Manier, D., and Hirst, W. (2006). The transformation of collective memories: studies of family recounting. *Mem. Cogn.* 34, 752–762.

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 373, 1–7.

Dunning, D. (2011). The dunning-kruger effect: on being ignorant of one's own ignorance. *Adv. Exp. Soc. Psychol.* 44, 247–296. doi: 10.1016/b978-0-12-385522-0.00005-6

Duque, J. F., Turner, R., Lewis, E. D., and Egan, G. (2010). Neuroanthropology: a humanistic science for the study of culture-brain nexus. *Soc. Cogn. Affect. Neurosci.* 5, 138–147. doi: 10.1093/scan/nsp024

Falk, E. B., and Bassett, D. S. (2017). Brain and social networks: fundamental building blocks of human experience. *Trends Cogn. Sci.* 21, 674–690. doi: 10.1016/j.tics.2017.06.009

Fernbach, P. M., Rogers, T., Fox, C. R., and Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychol. Sci.* 24, 939–946. doi: 10.1177/0956797612464058

Fisher, A. J., Medaglia, J. D., and Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6106–E6115.

Fisher, M., Goddu, M. K., and Keil, F. C. (2015). Searching for explanations: how the internet inflates estimates of internal knowledge. *J. Exp.l Psychol. Gen.* 144, 674–687. doi: 10.1037/xge0000070

Fisher, M., and Keil, F. C. (2014). The illusion of argument justification. *J. Exp. Psychol. Gen.* 143, 425–433. doi: 10.1037/a0032234

Flack, J. C. (2012). Multiple time-scales and the developmental dynamics of social systems. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1802–1810. doi: 10.1098/rstb.2011.0214

Frenda, S. J., Knowles, E. D., Saletan, W., and Loftus, E. F. (2013). False memories of fabricated political events. *J. Exp. Soc. Psychol.* 49, 280–286. doi: 10.1016/j.jesp.2012.10.013

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2019). *Cognitive Neuroscience: The Biology of the Mind*. New York, NY: Norton & Company.

Geana, M., Rabb, N., and Sloman, S. A. (2021). *Walking the Party Line: The Growing Role of Political Ideology in Shaping Health Behavior in the United States*. Manuscript under review

Goldberg, S. (2016). *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's the Meaning of Meaning: Twenty Years of Reflection on Hilary Putnam's the "Meaning of Meaning"*. London: Routledge.

Goldman, A. I. (1999). Knowledge in a social world. *Philos. Phenomenol. Res.* 64, 185–190.

Han, J. Y., and Federico, C. M. (2017). Conflict-framed news, self-categorization, and partisan polarization. *Mass Commun. Soc.* 20, 455–480. doi: 10.1080/15205436.2017.1292530

Hasson, U., and Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20150366. doi: 10.1098/rstb.2015.0366

Hawkins, C. B., and Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Pers. Soc. Psychol. Bull.* 38, 1437–1452. doi: 10.1177/0146167212452313

Hayek, F. A. (1945). The use of knowledge in society. *Am. Economic Rev.* 35, 519–530.

Hemmatian, B., Chan, S.-Y., and Sloman, S. A. (2019). "What gives a diagnostic label value? common use over informativeness," in *Poster at the Label entrenchment effect in explanation*, Chicago, IL.

Hemmatian, B., and Sloman, S. A. (2018). Community appeal: explanation without information. *J. Exp. Psychol. Gen.* 147, 1677–1712. doi: 10.1037/xge0000478

Henrich, J. (2015). Culture and social behavior. *Curr. Opin. Behav. Sci.* 3, 84–89.

Henrich, N., and Henrich, J. (2007). *Why Human Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.

Hirst, W., and Manier, D. (1996). "Social influences on remembering," in *Remembering the Past*, ed. D. Rubin (New York, NY: Cambridge University Press), 271–290. doi: 10.1017/CBO9780511527913.011

Holmes, R. M. (2020). *Cultural Psychology: Exploring Culture and Mind in Diverse Communities*. Oxford: Oxford University Press. doi: 10.1093/oso/9780199343805.001.0001

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1881.001.0001

Jacobson, G. C. (2010). Perception, memory, and partisan polarization on the Iraq war. *Polit. Sci. Q.* 125, 31–56. doi: 10.1002/j.1538-165x.2010.tb00667.x

Kahan, D. M., Jenkins-Smith, H., and Braman, D. (2011). Cultural cognition of scientific consensus. *J. Risk Res.* 14, 147–174. doi: 10.1080/13669877.2010.511246

Kominsky, J. F., and Keil, F. C. (2014). Overestimation of knowledge about word meanings: the "misplaced meaning". *Effect. Cogn. Sci.* 38, 1604–1633. doi: 10.1111/cogs.12122

Lawson, R. (2006). The science of cycology: failures to understand how everyday objects work. *Mem. Cognit.* 34, 1667–1675. doi: 10.3758/bf03195929

Marks, J., Copland, E., Loh, E., Sunstein, C. R., and Sharot, T. (2019). Epistemic spillovers: learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition* 188, 74–84. doi: 10.1016/j.cognition.2018.10.003

Marr, D. (1982). *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman.

Meade, M. L., Harris, C. B., Van Bergen, P., Sutton, J., and Barnier, A. J. (2018). *Collaborative Remembering*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198737865.001.0001

Mercier, H., and Sperber, D. (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., et al. (2002). Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* 16, 1159–1164. doi: 10.1006/nimg.2002.1150

Moreau, Q., and Dumas, G. (2021). Beyond correlation versus causation: multi-brain neuroscience needs explanation. *Trends Cogn. Sci.* 25, 542–543. doi: 10.1016/j.tics.2021.02.011

Nickerson, R. S. (1999). How we know – and sometimes misjudge - what others know: imputing one's own knowledge to others. *Psychol. Bull.* 125, 737–759. doi: 10.1037/0033-2909.125.6.737

Novembre, G., and Iannetti, G. (2020). Proving causality in hyperscanning: multibrain stimulation and other approaches: response to Moreau and Dumas. *Trends Cogn. Sci.* 25, 544–545. doi: 10.1016/j.tics.2021.03.013

Ober, J. (2008). *Democracy and Knowledge: Innovation and Learning in Classical Athens*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400828807

Parkinson, C., Kleinbaum, A. M., and Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nat. Hum. Behav.* 1:0072. doi: 10.1038/s41562-017-0072

Paulhus, D. L., Harms, P. D., Bruce, M. N., and Lysy, D. C. (2003). The over-claiming technique: measuring self-enhancement independent of ability. *J. Pers. Soc. Psychol.* 84, 890–904. doi: 10.1037/0022-3514.84.4.890

Pentland, A. S. (2012). The new science of building great teams. *Harv. Bus. Rev.* 90, 60–69.

Putnam, H. (1975). *Philosophical Papers*. New York, NY: Cambridge University Press.

Rabb, N., Fernbach, P., and Sloman, S. A. (2019). Individual representation in a community of knowledge. *Trends Cogn. Sci.* 23, 891–902. doi: 10.1016/j.tics.2019.07.011

Richerson, P. J., Boyd, R., and Henrich, J. (2010). Gene-culture coevolution in the age of genomics. *Proc. Natl. Acad. Sci. U. S. A.* 107(Suppl. 2), 8985–8992. doi: 10.1073/pnas.0914631107

Rilling, J. K., and Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annu. Rev. Psychol.* 62, 23–48. doi: 10.1146/annurev.psych.121208.131647

Rosch, E. (1978). "Principles of categorization," in *Cognition and Categorization*, eds E. Rosch, and B. B. Lloyd (Hillsdale, NJ: Erlbaum), 28–49.

Rozenblit, L., and Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* 26, 521–562. doi: 10.1207/s15516709cog2605_1

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., et al. (2013). Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414. doi: 10.1017/s0140525x12000660

Schmalzle, R., Hacker, F. E., Honey, C. J., and Hasson, U. (2015). Engaged listeners: shared neural processing of powerful political speeches. *Soc. Cogn. Affect. Neurosci.* 10, 1137–1143. doi: 10.1093/scan/nsu168

Searle, J. R. (2000). Consciousness. *Annu. Rev. Neurosci.* 23, 557–578. doi: 10.1146/annurev.neuro.23.1.557

Servick, K. (2020). In two-person MRI, brains socialize at close range. *Science* 367:133. doi: 10.1126/science.367.6474.133

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., and Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4687–E4696. doi: 10.1073/pnas.1323812111

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., et al. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* 7:12141. doi: 10.1038/ncomms12141

Sloman, S. A., and Fernbach, P. (2017). *The Knowledge Illusion : Why we Never Think Alone*. New York, NY: Riverhead Books.

Sloman, S. A., and Rabb, N. (2016). Your understanding is my understanding: evidence for a community of knowledge. *Psychol. Sci.* 27, 1451–1460. doi: 10.1177/0956797616662271

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proc Natl Acad Sci U S A* 107, 14425–14430. doi: 10.1073/pnas.1008662107

Theiner, G. (2013). Transactive Memory Systems: A Mechanistic Analysis of Emergent Group Memory. *Rev. Philos. Psychol.* 4, 65–89. doi: 10.1007/s13164-012-0128-x

Tomasello, M., and Carpenter, M. (2007). Shared intentionality. *Dev. Sci.* 10, 121–125.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–735. doi: 10.1017/s0140525x05000129

Toppi, J., Ciaramidaro, A., Vogel, P., Mattia, D., Babiloni, F., Siniatchkin, M., et al. (2015). "Graph theory in brain-to-brain connectivity: a simulation study and an application to an EEG hyperscanning experiment," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Burlington, MA, 2211–2214. doi: 10.1109/EMBC.2015.7318830

Turner, B. O., Santander, T., Paul, E. J., Barbey, A. K., and Miller, M. B. (2019). Reply to: fMRI replicability depends upon sufficient individual-level data. *Commun. Biol.* 2:129. doi: 10.1038/s42003-019-0379-5

Van Boven, L., Ehret, P. J., and Sherman, D. K. (2018). Psychological Barriers to bipartisan public support for climate policy. *Perspect. Psychol. Sci.* 13, 492–507. doi: 10.1177/1745691617748966

Vitriol, J. A., and Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *Eur. J. Soc. Psychol.* 48, 955–969. doi: 10.1002/ejsp.2504

Weatherford, J. (2005). *Genghis Khan and the Making of the Modern World. Broadway Books*. New York, NY: Crown and Three Rivers Press.

Wegner, D. M., Erber, R., and Raymond, P. (1991). Transactive memory in close relationships. *J. Pers. Soc. Psychol.* 61, 923–929. doi: 10.1037/0022-3514.61.6.923

Wilson, R. A. (2005). Collective memory, group minds, and the extended mind thesis. *Cogn. Process* 6, 227–236. doi: 10.1007/s10339-005-0012-z

Wilson, R. A., and Keil, F. (1998). The shadows and shallows of explanation. *Minds Mach.* 8, 137–159. doi: 10.1023/A:1008259020140

Wittgenstein, L. (1973). *Philosophical Investigations*, 3rd Edn. Oxford: Blackwell Publishing.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 686–688. doi: 10.1126/science.1193147

Zaller, J. (1992). *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511818691

Author Disclaimer: The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 05 November 2021
doi: 10.3389/fnsys.2021.772641

Understanding, Explanation, and Active Inference

Thomas Parr[1]* *and Giovanni Pezzulo*[2]

[1]*Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, London, United Kingdom,* [2]*Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy*

While machine learning techniques have been transformative in solving a range of problems, an important challenge is to understand why they arrive at the decisions they output. Some have argued that this necessitates augmenting machine intelligence with understanding such that, when queried, a machine is able to explain its behaviour (i.e., explainable AI). In this article, we address the issue of machine understanding from the perspective of active inference. This paradigm enables decision making based upon a model of how data are generated. The generative model contains those variables required to explain sensory data, and its inversion may be seen as an attempt to explain the causes of these data. Here we are interested in explanations of one's own actions. This implies a deep generative model that includes a model of the world, used to infer policies, and a higher-level model that attempts to predict which policies will be selected based upon a space of hypothetical (i.e., counterfactual) explanations—and which can subsequently be used to provide (retrospective) explanations about the policies pursued. We illustrate the construct validity of this notion of understanding in relation to human understanding by highlighting the similarities in computational architecture and the consequences of its dysfunction.

Keywords: active inference, explainable AI, insight, decision making, generative model, understanding

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Jakob Hohwy,
Monash University, Australia
Julian Kiverstein,
Academic Medical Center, Netherlands

***Correspondence:**
Thomas Parr
thomas.parr.12@ucl.ac.uk

Received: *08 September 2021*
Accepted: *15 October 2021*
Published: *05 November 2021*

Citation:
Parr T and Pezzulo G (2021) Understanding, Explanation, and Active Inference. Front. Syst. Neurosci. 15:772641. doi: 10.3389/fnsys.2021.772641

INTRODUCTION

How would we know whether a machine had understood why it chose to do what it did? Simplistically, we might expect that, when queried, it would be able to communicate an explanation for its actions. In this article, we take this to be our operational definition of *machine understanding* (Yufik, 2018). Based on this definition, we can break the problem down into two parts. The first is that a machine must be able to infer why it has taken the actions it has. The second is that it must be able to act to communicate this inference when queried. In thinking about the first—explaining behaviour—it is useful to think about how we go about explaining anything. In the philosophy of science, there is considerable debate about the notion of explanation (Craik, 1952; Bird, 1998; Psillos, 2002), which is beyond the scope of this article. Our use of the term is largely coherent with the idea of "inference to the best explanation" that is common in Bayesian treatments of perception (Helmholtz, 1866; Gregory, 1980) and in philosophy (Lipton, 2017) and proceeds as follows.

As scientists, we formulate a series of alternative hypothetical explanations. Each hypothesis entails different predictions about the data that we have measured. By comparing our predictions with those data, we assess which hypothesis is most congruent with our measurements. Translating this same process to explaining behaviour, the implication is that we need a space of hypotheses representing reasons[1] for behaviour, each of which predicts an alternative course of action. The process of explaining our actions[2]—i.e., having insight into our decisions—then becomes an inference problem. Given some observed sequence of choices, which explanations best fit those data?

This inferential perspective on decision-making is central to active inference (Friston et al., 2014; Parr et al., 2022), which frames perception and action as dual mechanisms that jointly improve our inferences about the causes of our sensory data. While perception is the optimisation of our beliefs to better fit the data we observe, action changes the world to better fit our beliefs. When the internal models required to draw these inferences are temporally deep (Friston et al., 2021), they must include the consequences of the sequential decisions we make while engaging with our environment. Active inference offers a set of prior beliefs about these decisions that represent explanations for behaviour. These explanations divide into three types (Da Costa et al., 2020). First, we select decisions whose sensory consequences cohere with the data anticipated under our model (Åström, 1965; Pezzulo et al., 2018). Second, our choices provide us with data that resolve our uncertainty about our environment (Mirza et al., 2018). Third, the context in which we find ourselves may bias us towards some actions and away from others (Pezzulo et al., 2013; Maisto et al., 2019). The first of these prompts us to head to a restaurant when our internal model predicts satiation when we feel hungry. The second leads us to survey the menu, to resolve our uncertainty about the food on offer. The third biases us towards ordering the same meal as on previous visits to the restaurant. Together, these account for exploitative (preference-seeking), explorative (curiosity-driven), and (context-sensitive) habitual behaviour. The last of these turns out to be particularly important in what follows, as it allows us to construct a narrative as to why we make the choices we do.

In what follows, we consider a simple, well-validated, task that incorporates both explorative, exploitative, and context-sensitive elements (Friston et al., 2015; Chen et al., 2020). It is based upon a T-maze paradigm, in which we start in the centre of the maze. In either the left or the right arm of the T-maze, there is a preferred (i.e., rewarding) stimulus whose position is initially unknown. In the final arm, there is a cue that indicates the location of the rewarding stimulus. To solve this maze and find the reward, we must decide whether to commit to one of the potentially rewarding arms or to seek out information about which is most likely to be profitable before exploiting this information. The twist here is that, after exposure to the maze, we follow up with a query. This takes the form of an instruction to explain either the first or the second move made. By communicating the reason for the action taken, the agent demonstrates a primitive form of insight into their own behaviour.

This touches upon questions about insight into our actions. This concept is important in many fields, ranging from metacognition (Fleming and Dolan, 2012) to cognitive neurology (Ballard et al., 1997; Fotopoulou, 2012) and psychiatry (David, 1990), where some syndromes are characterised by a patient exhibiting a lack of insight into their own behaviour. However, the term "insight" is often used to mean subtly different things and it is worth being clear upon the way in which we use the word here. Note that this is distinct from insight in the sense of the "aha moment"—where a different way of thinking about a problem leads to a clearer understanding of its solution (Kounios and Beeman, 2014; Friston et al., 2017a). In this article, we refer to insight of a different sort. Specifically, how do we come to understand the reasons for our own decision making? To the extent that veridicality is a useful concept here, insight can be regarded as a veridical inference about the causes of behaviour.

The hypothesis implicit in this article is that insight is confabulation, but that this confabulation may be constrained by sensory data to a greater or lesser extent. This provides a behavioural complement to the idea that perception is constrained hallucination (Paolucci, 2021). More precisely, both perception and explanation are inferences. In the extreme case that they are not constrained by data, we call them hallucination or confabulation, respectively. This perspective is endorsed by the philosophical position that, just as we must draw inferences about why other people behave the way they do, our explanations for our own behaviour are also inferred (Carruthers, 2009, 2011). However, we can go further than this. Interestingly, our retrospective (or confabulated) explanations are not innocuous but can change our beliefs about what we did and why. Specifically, hearing our own explanations provides further evidence for the policies we reported, which therefore become more plausible. This suggests an adaptive role for insight in improving our decision making, in addition to the benefits of being able to communicate explanations for behaviour to others.

In what follows, we briefly review the notion of a generative model and active inference. We then outline the specific generative model used throughout this and illustrate the behaviour that results from its solution through numerical simulations. Finally, we offer a summary of the results, in addition to a discussion of the relationship between the structure of these inferences in relation to the neuroanatomy of human cognition.

THE GENERATIVE MODEL

Under active inference, the generative model plays a central role in accounting for different sorts of behaviour. It is the implicit model used by a brain (or synthetic analogue) to explain the data presented by the environment. However, it is more than this. It also represents beliefs about how the world should be—from the perspective of some (biological or synthetic) creature (Bruineberg et al., 2016; Tschantz et al., 2020). This

[1]Note that not all hypotheses represent reasons for doing something, and the idea of reason does not follow directly from the scientific analogy. Central to this article is the idea that the hypotheses we are interested in can be translated into a verbal explanation that can be recognised as a reason for behaviour.

[2]Or the process of a machine or artificial agent explaining its actions.

means the generative model guides both a creature's perception and its actions. Formally, fulfilling these objectives requires scoring the quality of the model as an explanation of data and the quality of the data in relation to the model. The two qualities may be scored using a single objective function: the marginal likelihood or Bayesian model evidence. Simply put, the marginal likelihood scores the probability of observing some measured data given the model. That this depends upon both model and data implies it can be maximised either by modifying the model or by acting to change the data.

In practice, the marginal likelihood is often very difficult to compute. However, it can be approximated by a negative free energy functional (a.k.a., an evidence lower bound or ELBO). This free energy is constructed in relation to a variational (approximate posterior) distribution that maximises the free energy when it is as close as possible to the posterior probability of the hidden states in a model given measured data (Beal, 2003; Winn and Bishop, 2005; Dauwels, 2007). Some accounts of neuronal dynamics rest upon the idea that the activity in populations of neurons parameterises this variational density, and that the evolution of this activity ensures the alignment between the variational and exact posterior distributions (Friston and Kiebel, 2009; Bogacz, 2017; Parr et al., 2019; Da Costa et al., 2021). This means the role of a generative model in active inference is as follows. It determines the dynamics internal to some system (e.g., neuronal dynamics in the brain), and actions that result from these dynamics, *via* a free energy functional that approximates the marginal likelihood of the model. The maximisation of a marginal likelihood is sometimes characterised as "self-evidencing" (Hohwy, 2016).

We now turn to the specific generative model employed in this article. This is depicted in **Figure 1**. It is a deep temporal model (Friston et al., 2017b), in the sense that it evolves over two distinct timescales. Each level factorises into a set of factors [reminiscent of the idea of neuronal packets (Yufik and Sheridan, 1996)] that simplify the model—in the sense that we do not need to explicitly represent every possible combination of states (Friston and Buzsaki, 2016). At the faster (first) level, the model factorises into maze states and linguistic states. The former describes a T-maze in terms of two state factors (Friston et al., 2015). These are the agent's location in the maze, and the context—i.e., whether the reward is more likely to be in the left arm or the right arm. The location is controllable by the agent, in the sense that transitions between locations from one timestep to the next depend upon the choices it selects. These alternative transitions are indicated by the arrows in the location panel. Note that the left and right arms are absorbing states—meaning that once entered, the agent cannot leave these locations. In contrast, the context stays the same over time and cannot be changed through action. The allowable policies that the agent can select between are characterised in terms of sequences of actions (i.e., transitions). The first two moves across all the policies cover every possible combination of two moves (transitions to a given location), recalling that the absorbing states ensure that if the first move is to go to the left or right arm, the second move must be to stay there. The maze states predict two outcomes. The first is an exteroceptive outcome, that indicates where the agent is in the maze and, if at the cue location, whether the reward is most likely in the left or the right arm.

The second outcome modality pertains to the reward. Under active inference, there is nothing special about a reward modality: it is treated like any other observation. However, all outcome modalities can be assigned prior probability distributions that specify how likely we are to encounter the different outcomes in that modality. For instance, the generative model employed by a mouse might assign a relatively high prior probability to encounter cheese, in virtue of the fact that mice will act in such a way that they obtain cheese. For this reason, these prior probabilities can be regarded as prior preferences. A rewarding outcome is then simply a preferred or anticipated outcome. In other words, an outcome is rendered rewarding by the agent's anticipation of encountering it—and its actions to fulfil this expectation.

In our generative model, we include three levels of reward. The first is the attractive outcome (the reward) which is assigned a high relative prior probability. The second is an aversive outcome, which is assigned a low prior probability such that our agent believes it will act to avoid encountering it. The final outcome is a neutral outcome, with an intermediate prior preference. Depending upon the context, the attractive or aversive outcomes are encountered in the left and right arms of the maze, with the neutral outcome found elsewhere. The construction of the maze states is identical to that presented in previous articles, including (Friston et al., 2015, 2017; Chen et al., 2020).

The linguistic states are involved in determining the sentences that will be heard when the behaviour is queried or when responding to the query (i.e., the heard word and spoken word outcome modalities in **Figure 1**, respectively). As in previous applications of active inference to linguistic communication, these states factorise into syntactic structures and the semantics that can be expressed through this syntax (Friston et al., 2020). The syntactic states take the form of words and placeholder words associated with a set of transition probabilities—which determine which word (or placeholder) follows each other word. For instance, the word "Please" is followed by the word "explain." Depending upon the first word in the sequence, different syntactic structures appear. If we start with the word "Please", the syntax is consistent with a query. If starting with the word "I", it is an answer. In addition, there is a silent syntactic state associated with solving the maze. When the syntactic state is anything other than this silent state, the maze outcomes are set to be in the central location with a neutral reward. This precludes maze-solving (i.e., navigational) behaviour while the agent is attempting to explain its behaviour; and can be regarded as a form of sensory attenuation—as the maze states are functionally disconnected from their associated outcomes during the explanation. The semantic states are the words that can be slotted into the placeholders in the syntactic sequences to provide a meaningful sentence. The third semantic state doubles as the contextual state for the maze.

The slower (second) level deals with the narrative structure of the task, and the maintenance of the information required

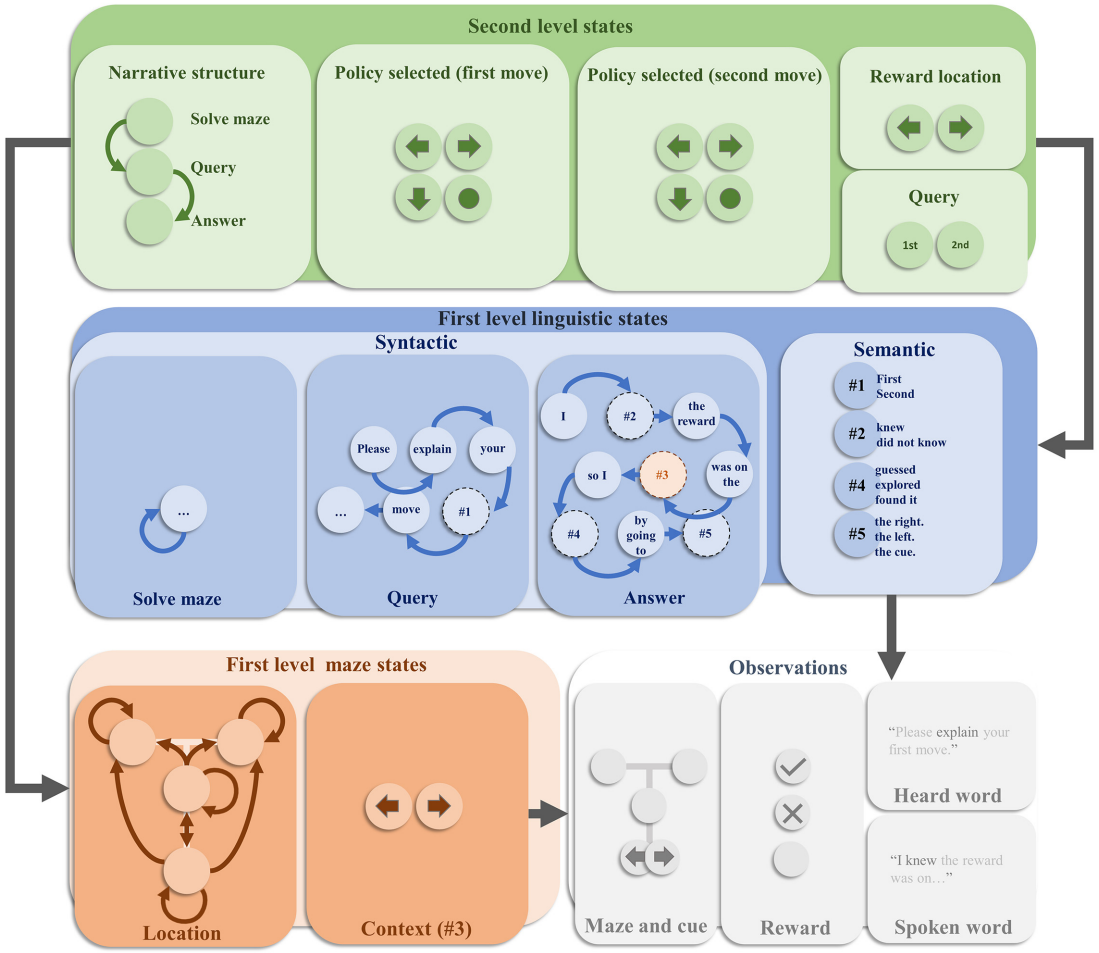


FIGURE 1 | The generative model. This schematic offers an overview of the internal model used by an agent to explain how hidden states conspire to generate observable outcomes. This figure is displayed in four main parts. These are the second level hidden states, the first level linguistic states, the first level maze states, and the observed outcomes. Each of these is further decomposed or factorised. The overall structure of the model means that second level states predict first level states. Although not shown here explicitly, the second level states additionally predict the policy (or trajectory) of the location states at the first level, providing a context sensitive bias for decision making. The first level states then combine to predict the observations. Arrows between states within each factor represent allowable transitions. In the absence of arrows, the assumption is that there are no dynamics associated with that state—i.e., it stays the same over time. Prior preferences are attributed to the outcomes such that the central location of the maze is mildly aversive. The reward outcome modality includes an attractive, aversive, and neutral outcome. Please see the main text for more detail.

for its solution. This includes a set of narrative states indicating whether the task is to solve the maze, listen to a query, or respond to that query. These are associated with a prior belief that the first thing to do is to solve the maze and that this is followed by the query and then the response. The narrative states predict the first syntactic state of the sequences at the first level. Specifically, the silent syntax is predicted when the maze should be solved, the syntax beginning "Please" when the query is offered, and the syntax beginning "I" when the answer is required. In addition, the policy is represented at the second level, decomposed into the first and second moves. Each combination of these predicts an alternative policy at the first level. The reward location state predicts the first level context, and the query state predicts whether the first level semantic state associated with the query syntax is "first" or "second"—i.e., whether the query is about the first or the second move. Combinations of these states predict different combinations of semantic states at the first level. For instance, when the narrative state is "answer", the query state is "first", and the first move state is a move to the cue location, the second semantic state is predicted to be 'did not know', the fourth semantic state is predicted to be "explored", and the fifth semantic state is predicted to be "the cue".

We will not unpack the details of the solution to this form of the generative model here, as they have been detailed in numerous other publications (Friston et al., 2017,b; Parr et al., 2019; Da Costa et al., 2020; Sajid et al., 2021). However, we provide a brief outline of the procedure. In short, the generative model outlined above can be formulated in terms of a joint probability distribution over the states ($s^{(i)}$) at each level (indicated by the superscript), the policy at the first level ($\pi^{(1)}$),

and the outcomes they generate (o). The marginal likelihood of this model can be approximated by a negative free energy functional (F) which can be recursively defined as follows:

$$\begin{aligned} F^{(2)}(o) &= D_{KL}\left[Q\left(s^{(2)}\right)||P\left(s^{(2)}\right)\right] \\ &+ \mathbb{E}_{Q(s^{(2)})}\left[F^{(1)}\left(o, s^{(2)}\right)\right] \\ F^{(1)}\left(o, s^{(2)}\right) &= D_{KL}\left[Q\left(\pi^{(1)}\right)||P\left(\pi^{(1)}|s^{(2)}\right)\right] \\ &+ \mathbb{E}_{Q(\pi^{(1)})}\left[F^{(1)}\left(o, \pi^{(1)}, s^{(2)}\right)\right] \\ F^{(1)}\left(o, \pi^{(1)}, s^{(2)}\right) &= D_{KL}\left[Q\left(s^{(1)}|\pi^{(1)}\right)||P\left(s^{(1)}|\pi^{(1)}, s^{(2)}\right)\right] \\ &- \mathbb{E}_{Q(s^{(1)}|\pi^{(1)})}\left[\ln P\left(o|s^{(1)}\right)\right] \end{aligned} \quad (1)$$

In Equation 1, the $\mathbb{E}$ symbol means "expectation" or average. The Q distributions are the variational distributions that approximate posterior probabilities and the symbol D_{KL} represents a Kullback-Leibler divergence—which quantifies how different two probability distributions are from one another. Beliefs about each set of states and policies in the model are computed as follows:

$$\begin{aligned} Q\left(s^{(2)}\right) &= \underset{Q(s^{(2)})}{\arg\min}\ F^{(2)}(o) \\ Q\left(\pi^{(1)}\right) &= \underset{Q(\pi^{(1)})}{\arg\min}\ \mathbb{E}_{Q(s^{(2)})}\left[F^{(1)}\left(o, s^{(2)}\right)\right] \\ Q\left(s^{(1)}|\pi^{(1)}\right) &= \underset{Q(s^{(1)}|\pi^{(1)})}{\arg\min}\ \mathbb{E}_{Q(s^{(2)})}\left[F^{(1)}\left(o, \pi^{(1)}, s^{(2)}\right)\right] \end{aligned} \quad (2)$$

The second line depends upon the empirical (conditional) prior probability for each policy. This is given as:

$$\begin{aligned} \ln P\left(\pi^{(1)}|s^{(2)}\right) &\propto \ln E\left(\pi^{(1)}|s^{(2)}\right) - G\left(\pi^{(1)}|s^{(2)}\right) \\ G\left(\pi^{(1)}|s^{(2)}\right) &\triangleq D_{\mathrm{KL}}\left[Q\left(o|\pi^{(1)}\right)||P(o|C)\right] \\ &+ \mathbb{E}_{Q(s^{(1)}|\pi^{(1)})}\left[H\left[P\left(o|s^{(1)}\right)\right]\right] \end{aligned} \quad (3)$$

Here, E is a function that acts as a prior weight or bias—conditioned upon the second level states, for the policies (Parr et al., 2021). The H in the second line is a Shannon entropy and C parameterises the preferred outcomes. The function G is referred to as expected free energy, and penalises those policies associated with large deviations from preferred outcomes, and policies in which the outcomes are uninformative about the hidden states.

This generative model permits two different types of action. These can be distinguished based upon how they influence the outcomes. The first sort of action influences the hidden states, which then cause changes in outcomes. Movement from one location to another in the maze falls under this category. In practice, these actions are chosen based upon the policy inferred to be the most probable *a posteriori*. The second sort of action directly influences the outcomes. This is the form of action involved in generating the linguistic outcomes (specifically, the spoken word). The latter are selected to minimise the free energy given current beliefs:

$$\begin{aligned} o_{\tau+1} &= \underset{o_{\tau+1}}{\arg\min}\ F^{(2)}\left(o_{\tau+1}|o_{t\leq\tau}\right) \\ &= \underset{o_{\tau+1}}{\arg\max}\ \mathbb{E}_{Q(s^{(1)},\pi^{(1)})}\left[\ln P\left(o_{\tau+1}|s^{(1)}_{\tau+1}\right)\right] \end{aligned} \quad (4)$$

This is in the same spirit as formulations of active inference in terms of predictive coding with reflexes. The idea is that by predicting the data we would anticipate given our beliefs, low-level reflexes of the sort found in the spinal cord or brainstem can correct deviations between our predictions and measurable data such that our predictions are fulfilled (Adams et al., 2013; Shipp et al., 2013). Having outlined the generative model, and the principles that underwrite its solution, we next turn to a series of numerical simulations that demonstrate some of the key behaviours of this model.

SIMULATIONS

In this section, we attempt to do three things. First, we illustrate the behaviour of an agent who relies upon the generative model outlined in the preceding section. We then attempt to offer some intuition as to the belief updating that underwrites this behaviour, and in doing so highlight the belief updating that occurs over multiple timescales in deep temporal models of this sort. In addition, we demonstrate the emergence of replay phenomena—of the sort that might be measured in the hippocampus of behaving rodents. Finally, we investigate what happens when we violate the assumptions of the generative model and the confabulatory explanations that result.

Figure 2 illustrates the behaviour and belief updating that occurs during the maze task, followed by the query presented to the agent and the response it offers. Two simulations are presented to show the answers given to two different queries, following the same behaviour. In both cases, the agent is initially uncertain about the context, as shown by the faint green circles in the left and right arms—indicating an equal posterior probability assigned to the reward being on the left or right. The agent starts in the central location and maintains veridical beliefs about its location throughout. At the second timestep, we see that the agent has elected to explore, seeking out the cue arm. On observing a cue indicating the right context, it updates its beliefs such that the reward is now anticipated in the right arm. At the third timestep, it has moved to the right arm, finding the reward there. When queried about the reasons for the first move, the agent sensibly replies that it did not know where the reward was (as we can verify from the plot of the maze at $t = 1$), so it explored by going to the cue location (as we can verify from the maze plot at $t = 2$). When queried about the second move, it replies that it did know where the reward was (again, verifiable from the maze plot at $t = 2$)—having already seen the cue by this point—and that it consequently went to find the reward in the right arm. This pair of simulations illustrates that the generative model is sufficient for the agent to infer the actions it has taken,

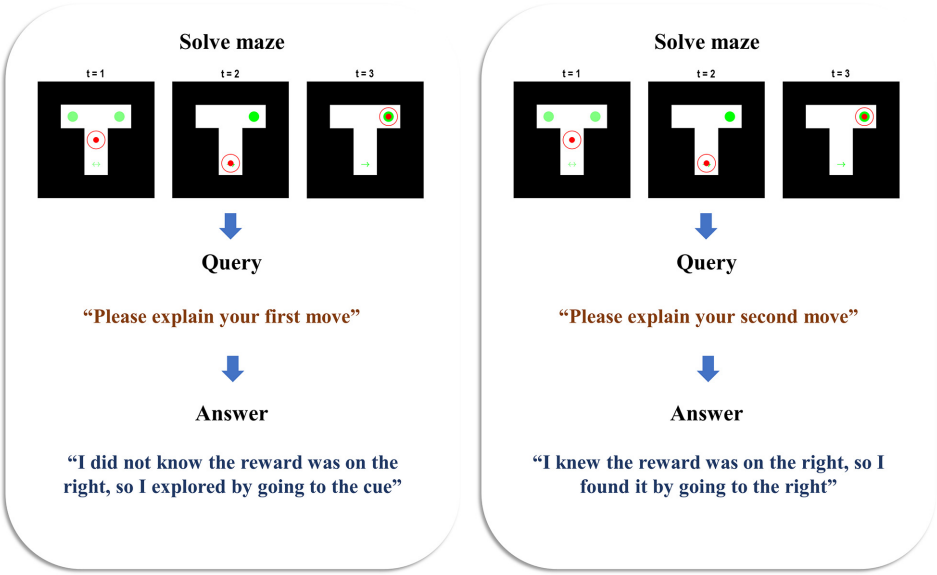


FIGURE 2 | Beliefs and behaviour. This figure reports the results of two simulations—one in which the first move is queried (left), and one in which the second move is queried (right). Each is divided into the three stages of the task. The first row of plots shows the beliefs and behaviour as the maze is solved during the first three time-steps. The small filled red circle shows the position of the agent in the maze. The larger unfilled circle shows the beliefs of the agent about its position. When this is red, this indicates a posterior probability of one of being in that location. As the probability decreases, the red fades to nothing—i.e., a posterior probability of zero of being in that location. In both simulations, the agent maintained confident and veridical beliefs about its location. The beliefs about the context (i.e., reward-left and reward-right) are shown in green, with the intensity of the green colour in the left and right arms (and of the arrows in the cue location) corresponds to the probability assigned to the associated context under the agent's posterior beliefs. On solving the maze, the agent moves on to the query stage, and the sentence presented to the agent is shown. Finally, the agent has the opportunity to answer the query, and the sentence it generates is shown below.

to come to a reasonable explanation of the motivations behind these actions, and to explain this when queried. In accordance with our definition in the introduction, this meets the criteria for a (simple) form of understanding.

To delve into the mechanisms by which this understanding is achieved, **Figure 3** details the beliefs held by the agent about the variables in the generative model throughout the simulation from the right of **Figure 2**. The grey dashed lines indicate the timesteps at the slower (second) level of the model, referred to hereafter as "epochs", and illustrate their alignment with the time-course of the faster) first level. During the first epoch, we see the first level beliefs (lower panel) being updated in accordance with the solution to the maze in **Figure 2**. The sequence here is reminiscent of the sequential activation of hippocampal place cells as rodents move through a series of locations (O'Keefe and Dostrovsky, 1971; Foster and Wilson, 2007). Inferences about the semantic states (i.e., the words #1 to #5 shown at the bottom of the bottom panel) remain uncertain during this time. Note the update in beliefs about the context (see "Right context" in the lower panel) on reaching the cue location, as the agent obtains the cue and goes from believing both contexts to be equally likely to believe that the right context is in play. The accompanying belief-updating for the policy (centre panel) shows that initially, the agent believes it will choose one of the many policies that start by going to the cue location, which correspond to the rows coloured in grey —consistent with the epistemic affordance associated with this location. On reaching the cue location, all uncertainty about the context is resolved, meaning the only remaining motivational drive is to obtain the cue. This prompts further belief updating about the policy, favouring the single policy in which the first move was to go to the cue and the second to the right arm. On enacting this policy and receiving the associated sensory input (i.e., observing itself going to the cue then right arm locations), the agent becomes confident that this is the policy it has pursued.

The inferred policy and context now allow for updating of beliefs about the first epoch at the second level. Practically, the updating of beliefs at each level happens asynchronously in this implementation, such that beliefs at the second level are updated following the updates at the first level. This asynchronous updating rests upon an adiabatic assumption, which means the two timescales in question may be treated under a mean-field assumption (i.e., approximately independently of one another). Consistent with the first level inferences, the second level beliefs over this epoch are updated such that the first and second moves inferred are consistent with the selected policy, and the reward location is consistent with the maze context. These beliefs are then used to provide empirical priors for the first level during the second epoch. Note the second epoch begins with a veridical belief about the policy selected and the maze context—ensuring these do not have to be re-inferred by the first level.

During the second epoch at the first level, the query is presented to the agent. Once the word "second" is heard, it is

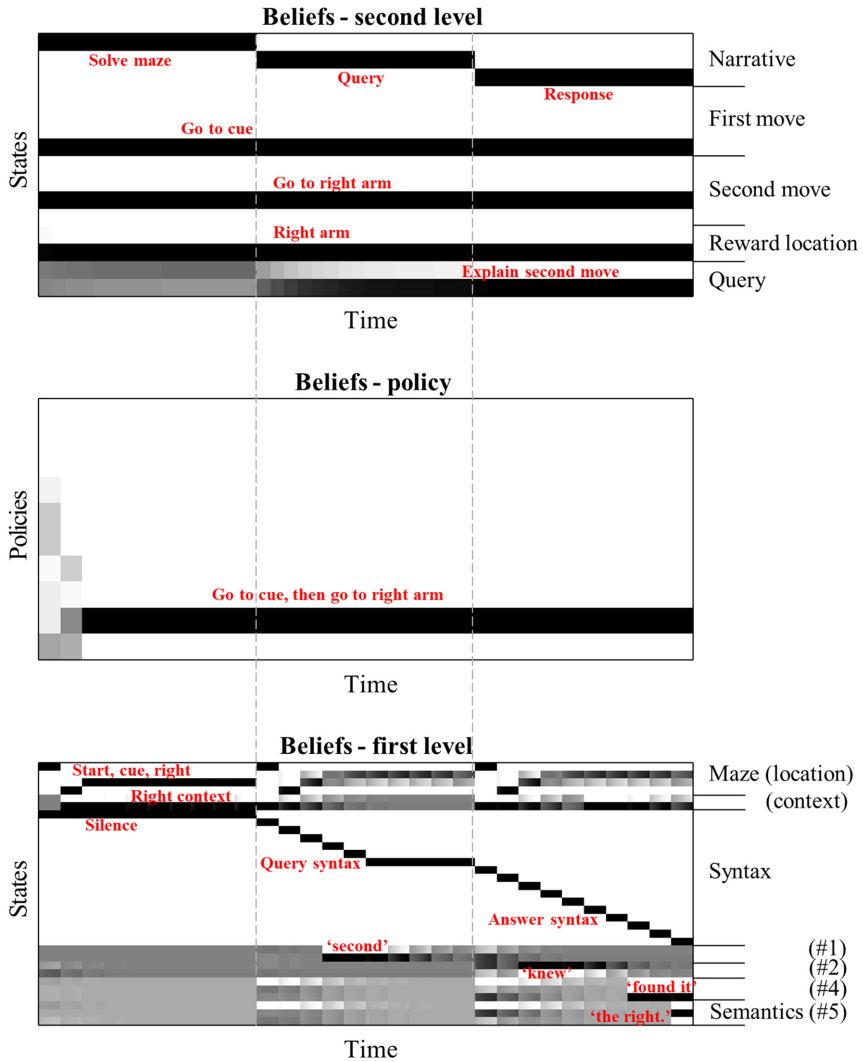


FIGURE 3 | Hierarchical belief updating. This figure shows the beliefs about states and policies over time in the temporally deep model. The main message of this figure is that this updating occurs over distinct timescales, with the first level states being updated much faster than those of the second level. The layout of these plots are as follows. Each row within each plot represents an alternative state or policy. The *x*-axis represents time; such that columns of the plot are discrete time steps. The shading of each cell in the state (or policy) x time arrays indicates the posterior probability assigned to that state (or policy) at that time. Black is a probability of one, white of zero, with intermediate shades representing intermediate probabilities. To avoid overcrowding, we have not labelled each row individually, but have annotated the states (and policies) that are inferred to be most likely in red. The vertical dashed lines indicate the alignment of the three epochs at the slowest (second level) with the inferences about the policies and states. Interpreted from a computational neuroscience point of view, each row of each plot can be regarded as a raster plot, indicating the aggregated firing rates of a distinct population of neurons.

able to update its belief about the first semantic state. At the end of the epoch, this is propagated to the second level, allowing for belief updating so that the query is inferred to be about the second move. This belief about the query is propagated through to the third epoch by the second level, again providing an empirical prior belief to the first level that the second query must now be answered. In addition, beliefs about the reward location and the moves selected at the second level combine with the belief about the query to provide prior beliefs about the semantic states. These beliefs lead the agent to generate the appropriate response to the question.

An interesting feature of the belief updating shown in **Figure 3** is the updates in beliefs about the maze location during the second and third epochs. Recall that, when the syntactic states are consistent with the query or answer, the maze states are decoupled from the associated outcomes—which are set at the central location and neutral reward. Despite this, the beliefs about the location in the maze during the first few timesteps of the second and third epoch appear to replay the beliefs that were held when solving the maze. **Figure 4** examines this more closely, by plotting the beliefs about the maze for the first three timesteps during each epoch. Note that, although the red dot

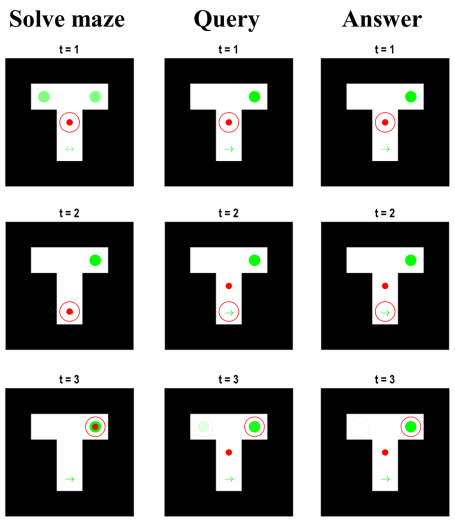


FIGURE 4 | Maze beliefs and replay. This figure reports the beliefs of the agent for the first three time-steps during each of the slower epochs at the narrative level. Each column of images displays a single epoch, with each row displaying the beliefs (and location) at each of the first three steps. The format of each image is the same as in **Figure 2**. These can be regarded as visual displays of the belief updates shown in **Figure 3**. The important things to note are: (i) that maze-solving epoch is identical to that of **Figure 2**, (ii) that the true location (i.e., red dot) is central in the query and answer epochs, (iii) that the context is known from the start in the latter two epochs, and (iv) that the beliefs about the location hidden states are consistent throughout all three epochs.

remains in the centre during the query and answer epochs, the red circle indicating the inferred location moves according to the same sequence as in the maze-solving epoch—however, the beliefs about the context are preserved from the end of the maze-solving. Replay of this sort has been identified physiologically in rodents in the same hippocampal cells that signal sequences of locations while behaving (Louie and Wilson, 2001; Foster, 2017; Pezzulo et al., 2017), hinting that the mechanisms that solve this generative model may also be at work in biological brains. That these mechanisms play a functional role is evidenced by the fact that interrupting hippocampal sharp-wave ripple activity (during which replays often occur) impairs memory-guided navigational choices (Jadhav et al., 2012). Interestingly, models built with the aim of simulating replay call upon a similar hierarchical structure in which the highest (narrative) level of the model involves an alternation between behaviour and replay sequences (Stoianov et al., 2021). Although these models focus more upon the role of replay in learning, our simulations suggest that such models can be interpreted, loosely, as if the synthetic agents are attempting to make sense of their previous actions during the replay sequences.

So why does replay occur when solving this model? The answer to this has two parts. In Bayesian statistics, the inference we draw depends upon a prior and a likelihood. In our model, both contribute to the development of replay. Recall that, while the query and answer syntactic states are in play, the maze outcomes are fixed. This means that no matter which actions the agent chooses, it will receive no sensory feedback coherent with those choices. This means the likelihood distribution is rendered uninformative and effectively uncouples the reality of the agent's position in the maze from the beliefs it has about this location. While this ensures the agent remains—for all practical purposes—fixed to the spot, it also liberates [or detaches (Gärdenfors, 2005; Pezzulo and Castelfranchi, 2007)] the inference process from the constraints of sensory input. As such, it can be seen as a form of sensory attenuation of the sort we might anticipate during dreaming (Windt et al., 2014), imagination (Villena-González et al., 2016; Kilteni et al., 2018), or episodic recall (Conway, 2001; Barron et al., 2020). This accounts for the role of the likelihood. However, freeing the agent from the constraints of sensation is not sufficient for replay. We also need a prior that assigns greater plausibility to the previous sequence of actions. This comes from the second level inference about the actions taken during the first epoch and their propagation to subsequent epochs as empirical priors. In other words, when sensory input is attenuated, a generative model simply recirculates prior information. In our example, this information pertains to the previous sequence of actions, but it could relate to other regularities learned during previous exposure to sensory data (Fiser et al., 2010; Buesing et al., 2011; Pezzulo et al., 2021). It is important to note that this construction was not designed to simulate replay. It is an emergent feature when beliefs about the policy must be propagated forwards in time (i.e., held in working memory) to help answer questions later.

Another feature of the belief updates from **Figure 2** that is worth unpacking is the increase in confidence about the context during the answer epoch. This seems counterintuitive, as the agent has had no new access to the maze outcomes. However, new data has arisen that prompts this increase in confidence—the agent has heard themself say that the reward was on the right. In other words, the agent is using its own answer about reward location as evidence about the context. To examine this further, **Figure 5** reports the results of a simulation in which we start with the query phase and provide two opportunities to answer the query. This means the maze is never solved (or, if it had been, no memory of the solution remains), but the agent is still asked about how they came to a (fictitious) solution, violating the assumptions of the generative model. In other words, it starts the query epoch with the same priors about policy and context as the agent in **Figures 3**, **4** has at the beginning of the maze-solving epoch. We see that, during the query epoch in **Figure 5**, the agent is uncertain about the state of the maze and the actions it took. It is confident that it started out in the first location and ascribes a slightly higher probability to being away from the central location by the third timestep, consistent with the fact that most plausible policies involve moving away from here. The probability of ending up in the left or right arm increases over time, as these are absorbing states. This in turn lends those policies leading to those states greater plausibility.

The beliefs about the left and right arms are similar during the first answer epoch (see the lower image of the centre column of **Figure 5**). This is because, by the third timestep of the first answer epoch, the agent has heard itself say that it did not know the reward location but has not yet heard its assertion that the reward was on the left. Taken together, the agent's first answer does not

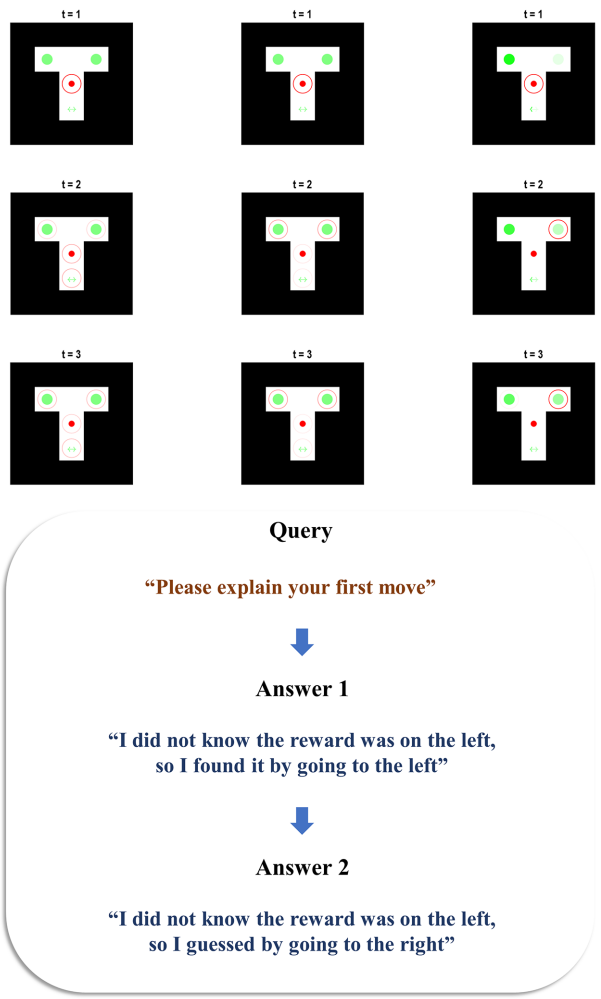


FIGURE 5 | Confabulation. This figure reports the behaviour of an agent who confabulates an answer as to why it pursued the course of action that it did, in virtue of never having pursued any course of action. It is given two opportunities to explain itself. The first answer given is nonsensical, as the generative model assumes that going straight to the reward.

make much sense. If going straight to the left on the first move, the agent should have known it was on the left in advance. This is even more puzzling when we note that the generative model takes a move that results in the reward location as evidence that the reward location was known. However, the apparent mismatch between not knowing it was on the left, but going straight there to find it anyway, is understandable when we consider that it is the second level of the model that enforces internal consistency in the story told by the agent. In our previous simulations (**Figure 2**), the agent already has a good idea as to which moves it made and the context of the maze by the time of the query epoch. In **Figure 5**, the agent is unable to formulate these beliefs until the first time it hears itself giving the explanation. However, by the second answer epoch, it has had a chance to synthesise what it has heard itself saying, and to revise this to an internally consistent explanation. Here, it has taken the fact that it did not know the reward location, and that it was on the left, and inferred that it must have guessed incorrectly given that it did not know the location. The result is the inference that it guessed at the reward location and got it wrong; a perfectly internally consistent, if confabulated, story.

DISCUSSION

This article was designed to address the problem of machine understanding and to show what this might look like using an active inferential approach in a simple example setting. The solution was based upon a deep temporal model, whose separation into two timescales allowed for a narrative overview of the task, and the propagation of information from one epoch to the next. The separation of timescales inherent to the model, and associated belief updating, in **Figure 3** is a generic feature of many deep temporal models. For example, in Friston et al. (2017b) a similar construction was used to simulate reading, where each word in a sentence provides information about the letters in the next word. In (Heins et al., 2020), a deep temporal model was employed for the purposes of a visual search paradigm, where each fixation point was associated with a dot-motion evidence accumulation task (Shadlen and Newsome, 1996). Similar approaches have been employed for working memory tasks (Parr and Friston, 2017), enabling the maintenance of information "in mind" throughout a delay period (Funahashi et al., 1989). These models have also found application in the modelling of emotions (Smith et al., 2019) and "affective inference" (Hesp et al., 2021). They have additionally been formulated through neural network models of the kind found in machine learning (Ueltzhöffer, 2018).

Probably the closest functional homologue to the process in this article was a deep temporal model of motor control (Parr et al., 2021), in which sequences of small movements were composed into longer trajectories *via* a higher (slower) level of the model. As in this article, this called upon the propagation not just of beliefs about states, but of beliefs about policies from one epoch to the next. One of the contributions of the modelling of motor control was to examine the consequences of a lesion to the connection between the two levels. Interestingly, lesions of the generative model for motor control produced a lack of coherence in movement trajectories that is formally analogous to the incoherent story confabulated during the first answer in **Figure 5** (later made consistent through the input from the second level).

The functional architecture of the homologous processes in the brain appears to involve the prefrontal cortices. Working memory is a good example of this, as the neural populations exhibiting persistent activity throughout delay-periods have been identified in the prefrontal cortex (Funahashi et al., 1989; Botvinick, 2008). However, these structures have also been linked directly to metacognition—the ability to assess one's own cognition—*via* lesion studies (Fleming et al., 2014). The frontal cortices interact directly—and *via* subcortical nuclei—with the temporal cortices (Kier et al., 2004; Blankenship et al., 2016; Rikhye et al., 2018), whose lateral surfaces are associated with

language (Price, 2000; Hutsler, 2003), and whose medial surfaces are associated with episodic memory and recall (Squire and Zola, 1998; Eichenbaum et al., 2012). They also share dense reciprocal connections with the basal ganglia (Naito and Kita, 1994)—the set of grey-matter nuclei most associated with the adjudication between alternative actions (Nambu, 2004). This hints at homology between the structure of the generative model in this article and the anatomy of the associated neuronal computation, providing a construct validation of our formulation of action understanding in relation to human understanding. It is also interesting to note that confabulatory pathologies in humans (Korsakoff, 1887) arise when the connectivity between the frontal, temporal, and subcortical nuclei are disrupted (Korsakoff, 1887; Benson et al., 1996; Turner et al., 2008), further endorsing the computational anatomy. Conceptually, starting from the query epoch in **Figure 5** may be analogous to a disconnection that precludes a memory of the maze-solving epoch from being propagated to the query epoch, such that we arrive at the query epoch as if it were the first epoch.

Given that the primary aim of this article was to address understanding of actions, it is interesting to note that some phenomena that feature in living machines (like us) emerge on solving this problem. The emergence of replay is of particular interest in the context of theories about the emergence of episodic memory. This form of memory has two defining features. The first is that it is declarative, in the sense that its contents can in principle be "declared" (Anderson, 1976; Squire and Zola, 1998). This contrasts with, for example, procedural memories. The second defining feature of episodic memory is that it is associated with a spatiotemporal context—in contrast to semantic memories of facts that may be divorced from such contexts. The replay phenomena shown in **Figures 3**, **4** meets both these criteria. We know it can be declared, as this is precisely what happens when the query is answered. It is spatiotemporal, in the sense that it is a memory of a sequence of locations in time. As such, one could view this as a simulation of a primitive sort of episodic memory. The reason this is interesting is that one explanation for episodic memory in biological creatures suggests that it developed alongside the ability to communicate past experiences (Mahr and Csibra, 2018). The simulations presented here lend some weight to these ideas, given that we set out to develop a model capable of explaining past actions, and found physiological hallmarks of episodic memory (i.e., replay) in the resulting belief updating. This is not in conflict with the conception of episodic memory as supporting a form of mental time-travel in the past and the future, enabling recall of the past and imagination of the future. As demonstrated in **Figure 5**, the agent is perfectly capable of using the same machinery for imagination of events that have not yet happened.

The central theme of this article is an inference about "what caused me to do that?" However, the status of Bayesian methods in establishing causation of this sort is controversial. The reason for this is that Bayes' theorem is symmetrical. It says that the product of a prior and likelihood is the product of a marginal likelihood and posterior probability. However, the labels "prior" and "marginal likelihood" can be swapped (provided "likelihood" and "posterior" are also swapped) without compromising the formal integrity of the theorem. This cautions against interpreting a conditional probability as a causal statement. This is less worrying in our context, as we know by construction that sensory data are caused by hidden states—i.e., we have implicitly built in a causal assumption to the model. However, the role of policies as causes of behaviour is a little more nuanced.

An influential formalism designed to address causality (Pearl, 2009, 2010) rests upon the idea of interventions. Under this formalism, an important notation is the "do" notation, in which $P(y|\,\mathrm{do}(x))$ is the distribution of y once x is fixed through some intervention. This breaks the symmetry of Bayes' theorem as, if x causes y, $P(y|\,\mathrm{do}(x))$ will be equal to $P(y|\,x)$, but $P(x|\,\mathrm{do}(y))$ will be equal to $P(x)$. The concept of intervention helps to contextualise the notion of a causal hierarchy—sometimes referred to as "Pearl's hierarchy" (Bareinboim et al., 2020). This hierarchy distinguishes between the three levels of the generative model. In ascending order, these are associational, interventional, and counterfactual. This provides a useful framework in which to situate the generative model outlined in this article. Given that the relationship between policies and the sequence of states is articulated in terms of conditional probabilities, our generative model must be at least at the associative level of Pearl's hierarchy. Implicitly, the interventional level criteria are also met, in that the inversion of the model employs a structured variational distribution (Dauwels, 2007) in which the marginals for the first level are evaluated as being independent of the second level states. This means the model is treated as if $P(s^{(1)}|\pi^{(1)})$ is equal to $P(s^{(1)}|\,\mathrm{do}(\pi^{(1)}))$ and $P(\pi^{(1)})$ is equal to $P(\pi^{(1)}|\,\mathrm{do}(s^{(1)}))$. However, it is worth noting that this applies only to the location states at the first level—the other states being conditionally independent of the policy given the second level states (i.e., the first level explanation is not caused by the policy pursued in an interventional sense, although there is an associational form of causality linking the two). In addition, the second level states do play a causative role, ensuring that the explanation at this level also causes the policy it attempts to account for. The third level of the Pearl hierarchy is more interesting from our perspective, given the emergence of a simple form of imagination as we saw in **Figure 5**. The criteria for counterfactual causation are met by noting that, initially, beliefs about all policies are evaluated for each policy. For each policy, this means there are a set of beliefs about states as if that policy were pursued. It is this counterfactual inference that facilitates the confabulation observed on asking for an explanation for a policy never pursued.

In this article, we elaborated on an operational notion of understanding as "inference to the best explanation" and described an active inference agent that is able to infer and communicate an explanation for its actions. However, the nature of understanding is a longstanding problem in philosophy—which we make no claims as to having solved. An interesting question is whether our agent (or more broadly, any artificial system) really understands anything. While addressing this question is clearly beyond the scope of this article, we hope that providing an example of an artificial system that appears to understand its actions helps advance the theoretical debate—and

assists in the identification of what is still missing from current operational definitions of understanding.

An interesting extension to this work would be to incorporate the response "I don't know" as an alternative to the explanations available to the agent. We have assumed this is unavailable to the agent in our simulations, as it is reasonable to assume that if we behave a particular way, we believe we know why we did. However, this prompts attempts at explanation despite not having engaged in the task. While it is interesting that these explanations enable the agent to convince itself of what has happened, we might anticipate that an agent could spare itself spurious explanations if able to infer that it is not sure of the answer. This might then point to the mechanisms for confabulation and loss of insight in psychopathology—framing it as a failure of inference about what is and is not known. However, this is not a straightforward problem to solve. This is evidenced by the (metacognitive) difficulties people have in assessing their own ability at performing even the simplest of tasks (Fleming et al., 2010; Fleming, 2021). Another interesting avenue would be to consider the role of two agents communicating with one another on task performance (Bahrami et al., 2012; Shea et al., 2014). For instance, it would be interesting to see whether, on receiving the explanation from an agent who has just completed the task, a second agent may perform the task more efficiently. Furthermore, the choice of question by the second agent may be more interesting, as they may wish to resolve uncertainty not just about the actions of the first agent, but of the structure of the task itself. For an example of this sort of diachronic inference, please see Friston et al. (2020). Diachronic inference refers to inferences drawn when two agents engage in a form of turn-taking, as is common in conversation, giving a periodic switching between speaking and listening. The current article dealt with only a single switch (from listening to speaking), which could usefully be expanded into a more extended conversation.

CONCLUSION

A key challenge for machine learning and artificial intelligence is to overcome the problem of understanding. While these approaches have been successful in making a range of decisions, the explanations for these decisions is often opaque. This article has sought to set out what a system capable of understanding and providing explanations for its decisions might look like. We took as our operational definition of understanding the ability to disambiguate between alternative hypotheses as to the reasons for behaving in a particular way and the ability to communicate the inferred reason for this behaviour, on being queried. To this aim, we constructed a generative model that predicts both behaviour and its (linguistic) explanation. This called upon a deep model that propagates information about choices through multiple epochs, enabling the presentation of a task (a simple T-maze), a query epoch, and an answer epoch. We demonstrated that inversion of this model under active inference allows for convincing explanations for the decisions made when solving the task. Interestingly, these explanations can also change our beliefs about what somebody did and why. Furthermore, biological phenomena such as replay emerge from this inversion—affording evidence for theories of episodic memory based upon a need to communicate past events. Finally, we saw that the pathologies of inference—on violation of the assumptions of the model—are similar to those seen in human behaviour in the context of some psychopathologies. The pathological explanations we encountered highlight that understanding can be thought of as constrained confabulation, but that it is constrained to a greater or lesser degree by the quality of the data used to form explanations.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found here: https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. All authors contributed to the article and approved the submitted version.

FUNDING

The Wellcome Centre for Human Neuroimaging is supported by core funding (203147/Z/16/Z). GP received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 945539 (Human Brain Project SGA3) and No. 952215 (TAILOR), and by the European Research Council under the Grant Agreement No. 820213 (ThinkAhead).

REFERENCES

Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643. doi: 10.1007/s00429-012-0475-5

Anderson, J. R. (1976). *Language, Memory and Thought*. UK: Lawrence Erlbaum.

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *J. Math. Anal. Appl.* 10, 174–205. doi: 10.1016/0022-247X(65)90154-X

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 367, 1350–1365. doi: 10.1098/rstb.2011.0420

Ballard, C., McKeith, I., Harrison, R., O'Brien, J., Thompson, P., Lowery, K., et al. (1997). A detailed phenomenological comparison of complex visual hallucinations in dementia with lewy bodies and Alzheimer's disease. *Int. Psychogeriatr.* 9, 381–388. doi: 10.1017/s1041610297004523

Bareinboim, E., Correa, D., Ibeling, D., and Icard, T. (2020). *On Pearl's Hierarchy and the Foundations of Causal Inference.* Columbia CausalAI Laboratory, Technical Report(R-60).

Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Prog. Neurobiol.* 192:101821. doi: 10.1016/j.pneurobio.2020.101821

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference.* United Kingdom: University of London.

Benson, D. F., Djenderedjian, A., Miller, B. L., Pachana, N. A., Chang, L., Itti, L., et al. (1996). Neural basis of confabulation. *Neurology* 46, 1239–1243. doi: 10.1212/wnl.46.5.1239

Bird, A. (1998). *Philosophy of Science.* London: McGill-Queen's University Press.

Blankenship, T. L., O'Neill, M., Deater-Deckard, K., Diana, R. A., and Bell, M. A. (2016). Frontotemporal functional connectivity and executive functions contribute to episodic memory performance. *Int. J. Psychophysiol.* 107, 72–82. doi: 10.1016/j.ijpsycho.2016.06.014

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *J. Math. Psychol.* 76, 198–211. doi: 10.1016/j.jmp.2015.11.003

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* 12, 201–208. doi: 10.1016/j.tics.2008.02.009

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 195, 2417–2444. doi: 10.1007/s11229-016-1239-1

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* 7:e1002211. doi: 10.1371/journal.pcbi.1002211

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behav. Brain Sci.* 32, 121–138. doi: 10.1017/S0140525X09000545

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press.

Chen, A. G., Benrimoh, D., Parr, T., and Friston, K. J. (2020). A bayesian account of generalist and specialist formation under the active inference framework. *Front. Artif. Intell.* 3:69. doi: 10.3389/frai.2020.00069

Conway, J. (2001). Sensory-perceptual episodic memory and its context: autobiographical memory. *Philos. Trans. R. S. Lond. B. Biol. Sci.* 356, 1375–1384. doi: 10.1098/rstb.2001.0940

Craik, K. J. W. (1952). *The Nature of Explanation.* Cambridge: Cambridge University Press.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *J. Math. Psychol.* 99:102447. doi: 10.1016/j.jmp.2020.102447

Da Costa, L., Parr, T., Sengupta, B., and Friston, K. (2021). Neural dynamics under active inference: plausibility and efficiency of information processing. *Entropy (Basel)* 23:454. doi: 10.3390/e23040454

Dauwels, J. (2007). "On variational message passing on factor graphs," in *2007 IEEE International Symposium on Information Theory*, (Nice: IEEE), 2546–2550.

David, A. S. (1990). Insight and psychosis. *Br. J. Psychiatry* 156, 798–808. doi: 10.1192/bjp.156.6.798

Eichenbaum, H., Sauvage, M., Fortin, N., Komorowski, R., and Lipton, P. (2012). Towards a functional organization of episodic memory in the medial temporal lobe. *Neurosci. Biobehav. Rev.* 36, 1597–1608. doi: 10.1016/j.neubiorev.2011.07.006

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* 14, 119–130. doi: 10.1016/j.tics.2010.01.003

Fleming, S. (2021). *Know Thyself: How the New Science of Self Awareness Gives Us the Edge.* London: John Murray Press.

Fleming, S. M., and Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philos. Trans. R. S. B. Biol. Sci.* 367, 1338–1349. doi: 10.1098/rstb.2011.0417

Fleming, S. M., Ryu, J., Golfinos, J. G., and Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 137, 2811–2822. doi: 10.1093/brain/awu221

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science* 329, 1541–1543. doi: 10.1126/science.1191883

Foster, D. J. (2017). Replay comes of age. *Annu. Rev. Neurosci.* 40, 581–602. doi: 10.1146/annurev-neuro-072116-031538

Foster, D. J., and Wilson, M. A. (2007). Hippocampal theta sequences. *Hippocampus* 17, 1093–1099. doi: 10.1002/hipo.20345

Fotopoulou, A. (2012). Illusions and delusions in anosognosia for hemiplegia: from motor predictions to prior beliefs. *Brain* 135, 1344–1346. doi: 10.1093/brain/aws094

Friston, K., and Buzsaki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cogn. Sci.* 20, 500–511. doi: 10.1016/j.tics.2016.05.001

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K., Da Costa, L., Hafner, D., Hesp, C., and Parr, T. (2021). Sophisticated inference. *Neural Comput.* 33, 713–763. doi: 10.1162/neco_a_01351

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017a). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017b). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402.

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., and Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* 118, 42–64.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 369:20130481. doi: 10.1098/rstb.2013.0481

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* 61, 331–349. doi: 10.1152/jn.1989.61.2.331

Gärdenfors, P. (2005). "The detachment of thought," in *The Mind as a Scientific Object: Between Brain and Culture*, eds C. E. Erneling and D. M. Johnson (New York, NY: Oxford University Press), 323–341.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 290, 181–197. doi: 10.1098/rstb.1980.0090

Heins, R. C., Mirza, M. B., Parr, T., Friston, K., Kagan, I., and Pooresmaeili, A. (2020). Deep active inference and scene construction. *Front. Artif. Intell.* 3:509354. doi: 10.3389/frai.2020.509354

Helmholtz, H. V. (1866). "Concerning the perceptions in general," in *Treatise on Physiological Optics*, ed J. P. C. Southall (New York: Dover), 1–37.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. D. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput.* 33, 398–446. doi: 10.1162/neco_a_01341

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hutsler, J. J. (2003). The specialized structure of human language cortex: pyramidal cell size asymmetries within auditory and language-associated regions of the temporal lobes. *Brain Lang.* 86, 226–242. doi: 10.1016/s0093-934x(02)00531-x

Jadhav, S. P., Kemere, C., German, P. W., and Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science* 336, 1454–1458. doi: 10.1126/science.1217230

Kier, E. L., Staib, L. H., Davis, L. M., and Bronen, R. A. (2004). MR imaging of the temporal stem: anatomic dissection tractography of the uncinate fasciculus, inferior occipitofrontal fasciculus and meyer's loop of the optic radiation. *Am. J. Neuroradiol.* 25, 677–691.

Kilteni, K., Andersson, B. J., Houborg, C., and Ehrsson, H. H. (2018). Motor imagery involves predicting the sensory consequences of the imagined movement. *Nat. Commun.* 9:1617. doi: 10.1038/s41467-018-03989-0

Korsakoff, S. (1887). Disturbance of psychic function in alcoholic paralysis and its relation to the disturbance of the psychic sphere in multiple neuritis of non-alcoholic origin. *Vestnik Psichiatrii* 4, 1–102.

Kounios, J., and Beeman, M. (2014). The cognitive neuroscience of insight. *Annu. Rev. Psychol.* 65, 71–93. doi: 10.1146/annurev-psych-010213-115154

Lipton, P. (2017). "Inference to the best explanation," in *A Companion to the Philosophy of Science*, ed. W. H. Newton-Smith (Oxford/Malden, MA: Blackwell), 184–193. doi: 10.1002/9781405164481.ch29

Louie, K., and Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* 29, 145–156. doi: 10.1016/s0896-6273(01)00186-6

Mahr, J. B., and Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behav. Brain Sci.* 41, 1–93. doi: 10.1017/S0140525X17000012

Maisto, D., Friston, K., and Pezzulo, G. (2019). Caching mechanisms for habit formation in active inference. *Neurocomputing* 359, 298–314. doi: 10.1016/j.neucom.2019.05.083

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS One* 13:e0190429. doi: 10.1371/journal.pone.0190429

Naito, A., and Kita, H. (1994). The cortico-pallidal projection in the rat: an anterograde tracing study with biotinylated dextran amine. *Brain Res.* 653, 251–257. doi: 10.1016/0006-8993(94)90397-2

Nambu, A. (2004). A new dynamic model of the cortico-basal ganglia loop. *Prog. Brain Res.* 143, 461–466. doi: 10.1016/S0079-6123(03)43043-4

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171–175. doi: 10.1016/0006-8993(71)90358-1

Paolucci, C. (2021). "Perception as controlled hallucination," in *Cognitive Semiotics: Integrating Signs, Minds, Meaning and Cognition*, ed C. Paolucci (Cham: Springer), 127–157.

Parr, T., and Friston, K. J. (2017). Working memory, attention and salience in active inference. *Sci. Rep.* 7:14678. doi: 10.1038/s41598-017-15249-0

Parr, T., Limanowski, J., Rawji, V., and Friston, K. (2021). The computational neurology of movement under active inference. *Brain* 144, 1799–1818. doi: 10.1093/brain/awab085

Parr, T., Markovic, D., Kiebel, S. J., and Friston, K. J. (2019). Neuronal message passing using mean-field, bethe and marginal approximations. *Sci. Rep.* 9:1889. doi: 10.1038/s41598-018-38246-3

Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain and Behavior.* Cambridge, MA: The MIT Press.

Pearl, J. (2009). Causal inference in statistics: an overview. *Statist. Surv.* 3, 96–146. doi: 10.1214/09-SS057

Pearl, J. (2010). An introduction to causal inference. *Int. J. Biostat.* 6:7. doi: 10.2202/1557-4679.1203

Pezzulo, G., and Castelfranchi, C. (2007). The symbol detachment problem. *Cogn. Process.* 8, 115–131. doi: 10.1007/s10339-007-0164-0

Pezzulo, G., Kemere, C., and van der Meer, M. A. A. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Ann. N. Y. Acad. Sci.* 1396, 144–165. doi: 10.1111/nyas.13329

Pezzulo, G., Rigoli, F., and Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front. Psychol.* 4:92. doi: 10.3389/fpsyg.2013.00092

Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* 22, 294–306. doi: 10.1016/j.tics.2018.01.009

Pezzulo, G., Zorzi, M., and Corbetta, M. (2021). The secret life of predictive brains: what's spontaneous activity for? *Trends Cogn. Sci.* 25, 730–743. doi: 10.1016/j.tics.2021.05.007

Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *J. Anat.* 197, 335–359. doi: 10.1046/j.1469-7580.2000.19730335.x

Psillos, S. (2002). *Causation and Explanation.* Netherlands: Acumen Publishing.

Rikhye, R. V., Gilra, A., and Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci.* 21, 1753–1763. doi: 10.1038/s41593-018-0269-z

Sajid, N., Ball, P. J., Parr, T., and Friston, K. J. (2021). Active inference: demystified and compared. *Neural Comput.* 33, 674–712. doi: 10.1162/neco_a_01357

Shadlen, M. N., and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proc. Natl. Acad. Sci. U S A* 93, 628–633. doi: 10.1073/pnas.93.2.628

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* 18, 186–193. doi: 10.1016/j.tics.2014.01.006

Shipp, S., Adams, R. A., and Friston, K. J. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends Neurosci.* 36, 706–716. doi: 10.1016/j.tins.2013.09.004

Smith, R., Parr, T., and Friston, K. J. (2019). Simulating emotions: an active inference model of emotional state inference and emotion concept learning. *Front. Psychol.* 10:2844. doi: 10.3389/fpsyg.2019.02844

Squire, L. R., and Zola, S. M. (1998). Episodic memory, semantic memory and amnesia. *Hippocampus* 8, 205–211. doi: 10.1002/(SICI)1098-1063(1998)8:3<205::AID-HIPO3>3.0.CO;2-I

Stoianov, I., Maisto, D., and Pezzulo, G. (2021). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *BioRxiv* [Preprint]. doi: 10.1101/2020.01.16.908889

Tschantz, A., Seth, A. K., and Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Comput. Biol.* 16:e1007805. doi: 10.1371/journal.pcbi.1007805

Turner, M. S., Cipolotti, L., Yousry, T. A., and Shallice, T. (2008). Confabulation: damage to a specific inferior medial prefrontal system. *Cortex* 44, 637–648. doi: 10.1016/j.cortex.2007.01.002

Ueltzhöffer, K. (2018). Deep active inference. *Biol. Cybern.* 112, 547–573. doi: 10.1007/s00422-018-0785-7

Villena-González, M., López, V., and Rodríguez, E. (2016). Orienting attention to visual or verbal/auditory imagery differentially impairs the processing of visual stimuli. *Neuroimage* 132, 71–78. doi: 10.1016/j.neuroimage.2016.02.013

Windt, J. M., Harkness, D. L., and Lenggenhager, B. (2014). Tickle me, I think I might be dreaming! Sensory attenuation, self-other distinction and predictive processing in lucid dreams. *Front. Hum. Neurosci.* 8:717. doi: 10.3389/fnhum.2014.00717

Winn, J., and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* 6, 661–694.

Yufik, Y. M. (2018). "GNOSTRON: a framework for human-like machine understanding," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, (Bengaluru, India: IEEE), 136–145.

Yufik, Y. M., and Sheridan, T. B. (1996). Virtual networks: new framework for operator modeling and interface optimization in complex supervisory Control systems. *Ann. Rev. Control* 20, 179–195.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

REVIEW
published: 18 November 2021
doi: 10.3389/fnsys.2021.735406

Understanding and Synergy: A Single Concept at Different Levels of Analysis?

Mark L. Latash[1,2]*

[1]Department of Kinesiology, The Pennsylvania State University, University Park, PA, United States, [2]Moscow Institute of Physics and Technology, Dolgoprudnyj, Russia

Biological systems differ from the inanimate world in their behaviors ranging from simple movements to coordinated purposeful actions by large groups of muscles, to perception of the world based on signals of different modalities, to cognitive acts, and to the role of self-imposed constraints such as laws of ethics. Respectively, depending on the behavior of interest, studies of biological objects based on laws of nature (physics) have to deal with different salient sets of variables and parameters. Understanding is a high-level concept, and its analysis has been linked to other high-level concepts such as "mental model" and "meaning". Attempts to analyze understanding based on laws of nature are an example of the top-down approach. Studies of the neural control of movements represent an opposite, bottom-up approach, which starts at the interface with classical physics of the inanimate world and operates with traditional concepts such as forces, coordinates, etc. There are common features shared by the two approaches. In particular, both assume organizations of large groups of elements into task-specific groups, which can be described with only a handful of salient variables. Both assume optimality criteria that allow the emergence of families of solutions to typical tasks. Both assume predictive processes reflected in anticipatory adjustments to actions (motor and non-motor). Both recognize the importance of generating dynamically stable solutions. The recent progress in studies of the neural control of movements has led to a theory of hierarchical control with spatial referent coordinates for the effectors. This theory, in combination with the uncontrolled manifold hypothesis, allows quantifying the stability of actions with respect to salient variables. This approach has been used in the analysis of motor learning, changes in movements with typical and atypical development and with aging, and impaired actions by patients with various neurological disorders. It has been developed to address issues of kinesthetic perception. There seems to be hope that the two counter-directional approaches will meet and result in a single theoretical scheme encompassing biological phenomena from figuring out the best next move in a chess position to activating motor units appropriate for implementing that move on the chessboard.

Keywords: referent coordinate, uncontrolled manifold, stability, motor equivalence, efference copy, iso-perceptual manifold

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Boris Prilutsky,
Georgia Institute of Technology, United States
Thomas Parr,
University College London, United Kingdom

***Correspondence:**
Mark L. Latash
mll11@psu.edu

Received: *02 July 2021*
Accepted: *01 November 2021*
Published: *18 November 2021*

Citation:
Latash ML (2021) Understanding and Synergy: A Single Concept at Different Levels of Analysis? Front. Syst. Neurosci. 15:735406. doi: 10.3389/fnsys.2021.735406

Abbreviations: f, function; F, force; L, length; λ, threshold of the stretch reflex; MU, motor unit; ORT, space orthogonal to the uncontrolled manifold; RC, referent coordinate; UCM, uncontrolled manifold; V_{UCM} and V_{ORT}, variance within the UCM and within ORT.

INTRODUCTION

Two terms, "understanding" (as used in cognitive neuroscience) and "synergy" (as used in movement neuroscience) seem to be closely related to each other. Indeed, *understanding* has been viewed as the discovery of co-variation between groups of relevant cognitive variables based on optimization, likely related to minimizing energy expenditure inside the system (Yufik, 2013, 2019). It has been also linked to one's ability to transform multiple lower-level concepts into a unified higher-level concept, meaning (Perlovsky, 2016). Understanding leads to overcoming the inertia of prior learning and enabling the construction of adequate responses under novel and unfamiliar circumstances (Yufik and Friston, 2016).

The word *synergy* has been used in the field of motor control with two implied meanings: Grouping numerous elements into stable groups to reduce the number of variables manipulated by the brain and co-varying group involvement with the purpose to ensure dynamical stability of actions in the unpredictable environment (reviewed in Bernstein, 1947; Latash, 2008, 2020a,b). Optimization ideas have been used broadly to account for the observed grouping of elements and their time evolution during typical actions (reviewed in Prilutsky and Zatsiorsky, 2002; Diedrichsen et al., 2010). So, both notions can be viewed as combinations of grouping plus co-variation plus optimization. Can they represent fundamentally similar neural mechanisms reflecting different stages of the evolutionary process, from *synergies* seen across numerous species to *understanding* claimed to be unique to the human species (Yufik, 2019)?

The contrast between the two notions becomes obvious if one considers typical spaces of variables where these notions are defined and applied: The spaces of mental models and meanings in studies of understanding vs. the spaces of variables from classical physics such as forces and coordinates (and their derivatives) in studies of synergies. The two notions and the corresponding spaces reflect two classes of approaches to neuroscience problems based on laws of nature: top-down and bottom-up. The former tries to describe aspects of cognition, including the one of understanding. It starts with accepting a set of axiomatic notions such as the mental model and meaning. The second starts from the interface with the inanimate world and operates with notions from classical physics, in particular classical mechanics. Of course, top and bottom are defined within this classification relatively arbitrarily. For example, one can start from classical physics and chemistry or even physics of elementary particles, and consider the simplest motor actions as examples of top-down analysis.

This article follows the bottom-up approach as compared to typical studies of cognition. It starts with trying to identify terms within the biology-specific adequate language (Gelfand, 1991; Gelfand and Latash, 1998), missing in the physics of inanimate nature. This leads to two important concepts, those of parametric control and spatial referent coordinates (RCs) originating from the classical equilibrium-point hypothesis (Feldman, 1966, 1986, 2015). Further, the concept of synergy is linked to arguably the most important feature of biological actions, their controlled task-specific stability. The ideas of synergic control and hierarchical control with spatial RCs are merged naturally (Latash, 2010, 2019, 2021a) leading to the possibility of ensuring dynamic stability of actions at levels ranging from groups of motor units to the whole body. This is an actively developed field with applications to such areas as motor learning, neurological disorders, and rehabilitation.

Further, we try to expand this approach to the field of perception. This development faces major problems with experimental verification because salient variables are not as readily measurable objectively. Nevertheless, there are promising recent theoretical and experimental studies suggesting the existence of percept-stabilizing synergies. At the end of the article, we return to the notion of understanding and try to link it to the stage of discovery during motor skill acquisition.

THE NEURAL CONTROL OF BIOLOGICAL ACTION

Bernstein was arguably the first to emphasize that the brain could not in principle prescribe such peripheral variables as forces and trajectories given the typical time delays associated with processing and conduction of neural signals, and time-varying changes in the external forces and intrinsic body states, which can never be perfectly predicted in advance (Bernstein, 1947; translation in Latash, 2020b). According to one of the influential theories of motor control, this problem is solved by using parametric control: biological movements are produced by changing parameters within the relations between actively produced forces and coordinates (reviewed in Feldman, 2015; Latash, 2019). In physical terms, these parameters have been associated with spatial referent coordinates for the involved effectors. Their physiological meaning is threshold for muscle activation associated with subthreshold depolarization of corresponding neuronal pools.

An alternative approach to problems of motor control and coordination has been developed assuming that the brain performs computations (addressed as "internal models", e.g., Wolpert et al., 1998; Kawato, 1999; Shadmehr and Wise, 2005) to plan, predict, and prescribe peripheral mechanical variables produced by muscles, joints, and other effectors. Major differences between this approach and the one following Bernstein's traditions have been reviewed earlier (Ostry and Feldman, 2003; Feldman and Latash, 2005; Feldman, 2015). The purpose of this article is not to contribute to these polemics but to follow Bernstein's definition and understanding of synergies and review recent studies exploring synergies at different levels and in different domains.

Within the classical equilibrium-point hypothesis for the control of a single muscle (Feldman, 1966, 1986), the salient parameter is the threshold (λ) of the stretch reflex expressed in units of muscle length and, simultaneously, representing subthreshold depolarization of the corresponding alpha-motoneuronal pool expressed in units salient for neurophysiological processes, millivolts. Changing λ can lead to various changes in peripheral variables such as muscle activation level, force (F), and length (L), depending on the external force field, in line with Bernstein's insight.

The idea of control with spatial RCs has been generalized to both multi-muscle systems that take part in typical functional actions and to intra-muscle subsystems. Whole-body actions, for example, pointing, are assumed to be controlled with a relatively low-dimensional RC specified at the level of task-relevant effectors, for example, a three-dimensional coordinate during typical arm reaching or pointing actions. Further, there is a sequence of few-to-many transformations leading to higher-dimensional RCs at hierarchically lower levels such as joints and muscles. This process is associated with apparent problems of redundancy because a small number of constraints are used to specify a large number of variables. As discussed later, the classical formulation of this problem (Bernstein, 1947; Turvey, 1990) is misleading and has to be replaced with the concept of *abundance* (Latash, 2012), which is not a source of the computational problem but an evolutionary advantageous design that ensures both stability of actions and their flexibility, i.e., adjustment to the changing external conditions.

Recently, the idea of control with RCs has been expanded in the opposite direction, i.e., inside the muscle (Madarshahian et al., 2021). Indeed, a number of muscles in the human body are viewed as combinations of compartments (Jeneson et al., 1990; Mariappan et al., 2010), i.e., groups of motor units united by both functional and anatomical criteria. Each compartment consists of numerous motor units, which may be viewed as the smallest unit of control. A motor unit is controlled by a single alpha-motoneuron and, as such, it obeys the law "all or none", which means that it can be recruited only as a whole. The contribution of a motor unit to muscle (or compartment) activation and mechanics can be varied by changing the frequency (f_{MU}) of action potential generation by the corresponding alpha-motoneuron.

Figure 1A illustrates the dependence between f_{MU} and the length of a group of muscle fibers forming the motor unit. It is characterized by the threshold of activation, λ_{MU} (motor units are typically recruited in an orderly fashion, from the smallest to the largest ones, Henneman et al., 1965) and the specific shape of the dependence of f_{MU} on muscle length. An increase in f_{MU} corresponds to an increase in the contribution of this particular motor unit to muscle force. Hence, the muscle $F(L)$ characteristic may be viewed as a superposition of motor unit $f_{MU}(L)$ characteristics (**Figure 1B**). Of course, expansion of the control with RC into spaces of muscle compartments and motor units is associated with even more glaring problems of redundancy or, if one accepts the concept of abundance, with even more opportunities to ensure dynamical stability of salient task-related performance variables.

Recently, the idea of control with RCs has been developed to account for a variety of phenomena including effects of motor adaptation to unusual force fields (Gribble and Ostry, 2000), motor learning (Turpin et al., 2016), neuronal population coding of control variables by the brain (Feldman, 2019), agonist-antagonist coactivation (Latash, 2018a), perceptual errors (Latash, 2018b), and certain types of neurological disorders including spasticity (Jobin and Levin, 2000; Mullick et al., 2013). This approach is based on the solid foundation of experimental findings in studies ranging from those involving animal preparations (Feldman and Orlovsky, 1972; Hoffer and Andreassen, 1981) to healthy humans (Feldman, 1966; Schmidt and McGown, 1980; Latash and Gottlieb, 1990; Latash, 1992).

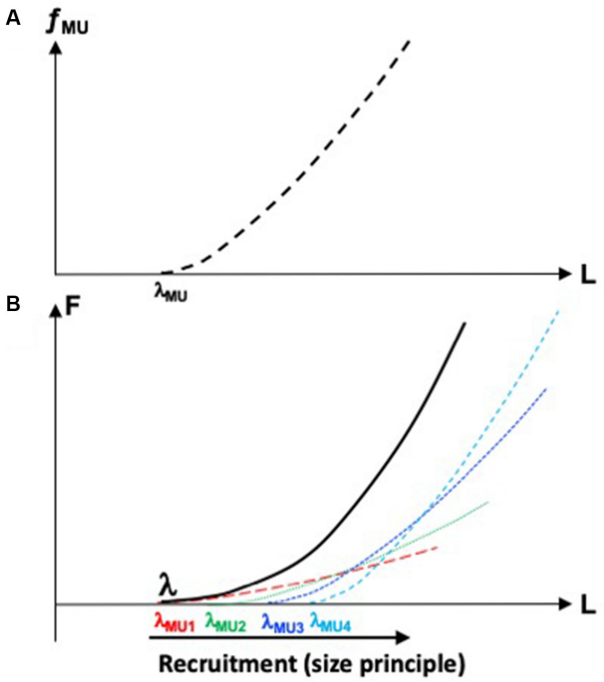


FIGURE 1 | (A) The dependence between frequency of firing of a motor unit, f_{MU}, and length of the group of muscle fibers forming this motor unit. Its threshold of activation is λ_{MU}. **(B)** The muscle force-length, $F(L)$ characteristic (solid line) may be viewed as a superposition of motor unit characteristics (dashed lines). The motor units are recruited in an orderly fashion, from smaller to larger ones.

CONTROLLED STABILITY OF ACTION

The concept of *synergy* in movement studies has been used at least since the XIXth century as a synonym of the word *coordination*; respectively, *asynergia* and *dyssynergia* have been used as synonyms of impaired coordination (Babinski, 1899). Bernstein incorporated this concept into his multi-level hierarchical scheme for the control of movements as the second from the bottom level. Its full name was "The level of synergies and patterns or the thalamo-pallidar level" emphasizing the importance of the loops through the basal ganglia, an insight supported by recent studies (reviewed in Latash and Huang, 2015). According to Bernstein, the level of synergies serves two main functions: (1) organizing numerous elements into groups; and (2) ensuring the dynamical stability of movements.

The former function of synergies is directly related to the famous problem of motor redundancy (Bernstein, 1947, 1967). Bernstein was arguably the first to pay attention to the fact that each natural movement involves numerous elements at multiple levels of analysis, kinetic, kinematic, muscle activation, etc. The number of elements is larger than the number

of constraints associated with typical tasks and, therefore, an infinite number of solutions exist. In his main book, Bernstein (1947) was ambiguous with respect to this problem. In different sections, he emphasized both the elimination of redundant degrees-of-freedom considered as the main problem of motor control and benefits of having extra degrees-of-freedom. How does the central nervous system select specific solutions observed during movements? Bernstein's expression "elimination of redundant degrees-of-freedom" as the method of finding unique solutions for typical problems of motor redundancy dominated the field until recently. In fact, the problem of motor redundancy has another component: Even for a single element, movement from an initial to a final state can proceed along an infinite number of trajectories. How does the central nervous system select specific trajectories from this set? So, there is a problem of state redundancy and a problem of trajectory redundancy. During natural movements, both problems coexist.

Arguably, the most commonly used method to solve such problems has been optimization formulated as search for a minimum (or maximum) of a cost function in different spaces of variables, mechanical, neurophysiological, and psychological (reviewed in Seif-Naraghi and Winters, 1990; Prilutsky and Zatsiorsky, 2002). Recently, methods of optimal feedback control have been used to find solutions for such problems (Todorov and Jordan, 2002; Diedrichsen et al., 2010). There are two obvious problems with most such methods. First, they assume that the neural controller computes cost function values, typically based on performance variables, over movement time prior to movement initiation, i.e., that movement time is known in advance and time profiles of the relevant variables can be accurately predicted over the future movement. Second, the choice of the cost function is usually rather arbitrary, reflecting personal theoretical preferences.

The ill-posed nature of the problem of motor redundancy can be illustrated with the example of excessive muscle co-activation seen at early stages of skill acquisition (Bernstein, 1947). Bernstein viewed this phenomenon as an attempt to mitigate the problem of redundancy by limiting the kinematic space of possible movements. This may be true if the problem is considered at the level of joint kinematics. However, co-activation obviously makes the problem worse at the level of muscle activation and motor unit recruitment. This example suggests that, before the problem is solved, it has to be clearly formulated at the level of neural control variables, such as RCs, not peripheral mechanical variables.

Recently, the problem of motor redundancy has been reformulated as the *principle of abundance* (Gelfand and Latash, 1998; Latash, 2012). This reformulation emphasizes the importance of variability in both neural and motor processes and postulates that the brain facilitates "good enough" solutions (Loeb, 2012; Akulin et al., 2019) and uses the abundance of elements to ensure desired dynamical stability of those solutions with respect to salient performance variables. The idea of abundance follows naturally the classical Bernstein's study of hammering by professional blacksmiths (Bernstein, 1930) where he showed that the trajectory of the tip of the hammer showed less inter-trial variability compared to the trajectories of individual joints. The importance of motor variability has also been illustrated by pathologies characterized by unusually low variability (e.g., low postural sway in advanced-stage Parkinson's disease, Horak et al., 1992) and the links between low variability and incidence of chronic pain in healthy persons (Madeleine et al., 2008; Madeleine and Madsen, 2009).

The principle of abundance fits well the aforementioned definition of the level of synergies in the multi-level hierarchical control scheme by Bernstein (1947) and Latash (2020a), in particular its assumed role in ensuring dynamical stability of actions. This approach is tightly linked to the concept of *uncontrolled manifold* (UCM; Schöner, 1995; Scholz and Schöner, 1999). According to the UCM-hypothesis, the central nervous systems acts in multi-dimensional spaces of elemental variables and structures variance in those spaces to allow relatively large variance along a subspace where a salient performance variable does not change (the UCM for that variable) while minimizing variance leading to changes in that variable, i.e., in the orthogonal to the UCM space (ORT space).

Figure 2 illustrates the UCM concept for the task of producing constant total force (F_{TOT}) while pressing with two independent effectors, e.g., two fingers. The inter-trial data cloud is expected to form an ellipse elongated along the UCM. Quantifying variance per dimension within the UCM and within the ORT is expected to produce an inequality $V_{UCM} > V_{ORT}$ if indeed the central nervous system stabilizes the potentially important performance variable (F_{TOT}, in this example) at the expense of other variables produced by the same set of effectors. If the subject in this experiment is asked to produce a different force magnitude, the UCM shifts, but the location and shape of the data cloud are expected to be robust (as illustrated for three F_{TOT} magnitudes in **Figure 2**). It has been suggested that the location of the center of the inter-trial cloud may reflect an optimization principle, whereas the shape of the cloud reflects the stability of the performance variable (Park et al., 2010). Assuming that there exists a single optimal solution and any deviations from this solution incur extra costs, large V_{UCM} (reflecting high stability) implies large deviations from the center of the data point distributions, i.e., large violations of the optimality principle.

Large magnitudes of V_{UCM} are reflections of low stability along the UCM, which is functionally important. Indeed, large V_{UCM} allows performing secondary tasks with the same set of elements without negative interference with the original task. In addition, low stability along the UCM channels effects of unexpected perturbations into the UCM thus protecting the salient variable from such perturbations. For example, imagine walking along the beach while carrying in the dominant hand a mug filled with hot coffee. At the level of kinematics, vertical mug orientation is a salient performance variable, which gets contributions from numerous kinematic variables—joint angles along the body and the arm. During walking, unexpected perturbations emerge frequently, e.g., when stepping on a pebble, unexpected surface, etc. A multi-joint synergy stabilizing the

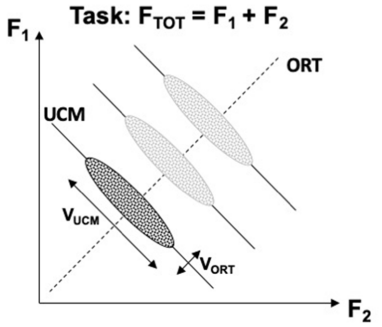


FIGURE 2 | An illustration of the UCM concept for the task of producing constant total force (F_{TOT}) while pressing with two fingers. A cloud of data points across trials is shown with an ellipse. Note that variance along the solution space (uncontrolled manifold, UCM, solid lines) is larger than orthogonal to the UCM (ORT, dashed line), $V_{UCM} > V_{ORT}$. Producing different force magnitudes is associated with UCM shifts, while the relative location and shape of the data clouds remain about the same (lighter images). V_{UCM}, variance within the UCM; V_{ORT}, variance within the ORT.

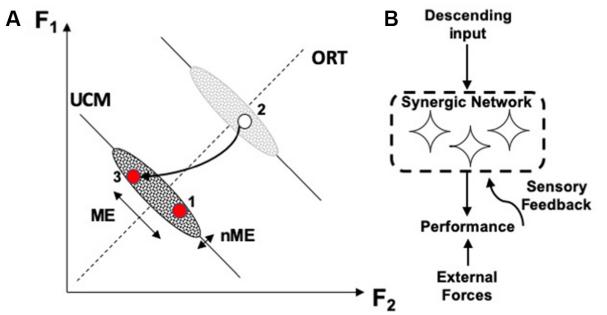


FIGURE 3 | (A) A perturbation leads to a deviation of the system from an initial point 1 to point 2. During quick corrective action (to point 3), motor equivalent (ME) deviations along the corresponding UCM are larger than non-ME deviations. **(B)** Large ME deviations are expected if corrective signals serve as perturbations into the neural network ensuring the corresponding synergy.

mug orientation helps channel the kinematic effects of such perturbations into the respective UCM. You can lean and pick up a shell without spilling the coffee, which requires using a subset of joints of the body; this can be done by limiting joint rotations to the UCM. Clinical studies have confirmed the importance of high V_{UCM} magnitudes by showing that low indices of stability seen in certain groups of neurological patients are associated primarily with low magnitudes of V_{UCM}, not with large magnitudes of V_{ORT} (Falaki et al., 2017; Jo et al., 2017).

A number of schemes have been suggested leading to the typical structure of variance for stabilized performance variables ($V_{UCM} > V_{ORT}$). These include short-latency feedback loops within the central nervous system, somewhat similar to the classical system of recurrent inhibitions, as well as feedback projections from peripheral sensory endings (Latash et al., 2005; Martin et al., 2009). Similar clouds of data points elongated along the solutions space have been reported in modeling studies based on the minimal intervention principle (Todorov and Jordan, 2002) and implemented using optimal feedback control schemes (reviewed in Diedrichsen et al., 2010). Within those schemes, deviations in spaces of elemental variables are corrected by the central nervous system only if they introduce errors into salient performance variables.

The different stability along the UCM and along ORT leads to a particular signature of the phenomenon of motor equivalence. If a person is instructed to correct an ongoing action in cases of perturbations affecting a salient performance variable, corrections show very large motor equivalent components, i.e., deviations along the corresponding UCM (**Figure 3**; Mattos et al., 2011, 2015). In other words, deviations of elemental variables during the corrections show large components that do not correct anything, i.e., they are wasteful from the point of view of energy expenditure. Such large motor equivalent deviations are expected if corrective signals generated by the brain are seen as inputs (perturbations) into a neural network forming the corresponding synergy (**Figure 3**). Studies of motor equivalent and non-motor equivalent deviations have confirmed their relationship to the V_{UCM} and V_{ORT} indices expected from statistics of folded distributions (Falaki et al., 2017).

Recently, the notion of performance-stabilizing synergies has been developed for spaces of hypothetical control variables, i.e., RCs at different levels of the presumed control hierarchy (Reschechtko and Latash, 2017, 2018; Latash, 2021a). Indeed, the abundance of RCs at any control level allows (but does not dictate!) synergies stabilizing performance. Such synergies have been confirmed in multi-finger force production tasks (Ambike et al., 2016a,b, 2018).

Important findings in studies of motor synergies include the following (reviewed in Latash, 2008, 2019)

- The central nervous system can use a set of elemental variables to stabilize various performance variables in a task-specific manner;
- Synergies can be attenuated in anticipation of an action that requires a quick change in the salient performance variable. These phenomena have been addressed as anticipatory synergy adjustments;
- Unintentional drifts in performance are associated with loss of stability, which can be quantified in spaces of mechanical elemental variables and control variables; and
- Controlled stability suffers with advanced age, atypical development, and a range of neurological disorders. It can be improved with specialized training.

THE ORIGIN OF STABLE AND ILLUSORY PERCEPTS

Perception of one's own body configuration, movements, and forces at the interface with the environment has been traditionally addressed as kinesthetic perception. Kinesthetic perception can be viewed as the process of measurement of salient variables and reporting them to oneself or others. The importance of both the efferent (motor related) and the afferent (sensory, generated at the periphery) signals for

kinesthetic perception has been accepted for a long time, at least from the middle of the last century when Von Holst and Mittelstaedt (1950/1973) introduced the notion of *efference copy*, close in spirit to the notion of *corollary discharge* (Sperry, 1950). In the original formulation, the concept of efference copy was associated with a copy of signals sent by alpha-motoneurons to muscles. This signal was used to predict changes in afferent signals from proprioceptors induced by the future movement (so-called, *reafference*). Reafference was expected to interact with efference copy and produce reflex changes in movements only if it differed from the efference copy-based prediction. This understanding of efference copy has been criticized recently (Feldman, 2009, 2016) because it cannot explain how muscles can be relaxed after a movement to a new posture. Indeed, if muscles are relaxed efference copy is the same (zero) in both states, and any changes in afferent signals cannot be predicted based on efference copy changes. Hence, they have to produce reflex muscle activation in contrast to everyday observations.

In more recent studies, the role of the efferent process in perception has been associated with specifying a reference point (RC, see earlier). Indeed, to measure a physical variable, one has to have a reference point (from where to measure) and a tool (e.g., a ruler to measure distance). The efferent process has been assumed to supply the former component, and signals from peripheral receptors—the latter component (reviewed in Feldman, 2015, 2016; Latash, 2019, 2021b). **Figure 4** illustrates the process of perceiving muscle length and force. Command to the muscle specifies the threshold of its stretch reflex (λ), which plays the role of RC. Many sensory signals show non-zero levels of activity when muscle length is shorter than λ and increase their activity level with deviation from λ along the force-length characteristic. These involve signals from length-sensitive and force-sensitive receptors as well as signals generated by the alpha-motoneurons innervating the muscle. Taken together, these signals form an abundant set, which may be viewed as the basis for stable kinesthetic percepts.

Imagine that you press with a hand against a stop such that no movement occurs. During changes in the pressing force, we have a veridical, undisturbed perception of steady posture. Where does this percept come from? Indeed, all signals from relevant peripheral receptors change. Signals from muscle spindles change with unavoidable changes in muscle fiber length (coupled to tendon length changes, such that the "tendon plus muscle" complex stays at the same length) and also changes in the activity of gamma-motoneurons, which change the sensitivity of spindle endings. Note that gamma-motoneurons change their activation level in parallel to the signals from alpha-motoneurons. There will be obvious changes in signals from force-sensitive Golgi tendon organs and from articular receptors, which are sensitive to the articular capsule tension. All the efferent signals will change as well. There seem to be no signals that are kept unchanged to correspond to the undisturbed perception of arm configuration. This observation has been interpreted as a reflection of all the signals, afferent and efferent, being constrained to a sub-space in the combined multi-dimensional afferent-efferent space—the iso-perceptual manifold (Latash, 2018b).

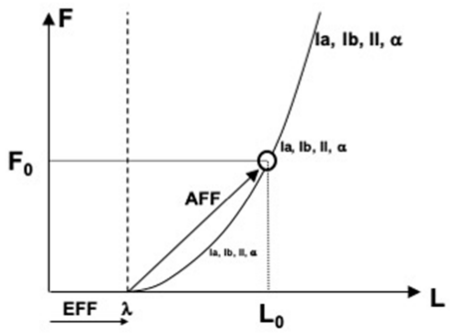


FIGURE 4 | An illustration of perceiving muscle length (L) and force (F). Command to the muscle specifies the threshold of its stretch reflex (λ), which plays the role of referent coordinate. Many sensory and motor signals increase with deviation from λ along the force-length characteristic. Any of these signals may serve as afferent components of perceiving both force and length, F_0 and L_0.

A cartoon illustration of the iso-perceptual manifold in a three-dimensional space is shown in **Figure 5**. One coordinate corresponds to an efferent signal (RC), and the other two—to two afferent signals (A_1 and A_2). Note that variations of all signals are possible within the iso-perceptual manifold leading to the undisturbed perception of the salient variable. Such motion can be termed *perceptually-equivalent*, similarly to the motor equivalent motion described earlier. When the signals go outside the iso-perceptual manifold, perception of a change in the respective variable is reported, even if it is kept unchanged. The concept of the iso-perceptual manifold can be viewed as a definition of a stable kinesthetic percept. Indeed, there is no other definition addressing perceptual stability, which is a functionally very important feature of perception, crucial in the evolutionary process.

The iso-perceptual manifold concept implies that accurate perception of a functionally important variable can be associated with variable efferent and afferent signals to and from the involved elements. As a result, perception of variables produced by the elements may be less accurate when they participate in a multi-element action compared to their perception in single-element actions and to the perception of a variable produced by all the elements together. This prediction has been confirmed experimentally showing that perception of finger force is more precise and less variable during single-finger force production tasks as compared to multi-finger tasks (Cuadra and Latash, 2019; Cuadra et al., 2021b).

The described scheme can account for kinesthetic illusions, in particular those induced by muscle vibration (Goodwin et al., 1972; Roll and Vedel, 1982; Lackner and Taublieb, 1984), a powerful stimulus for signals from velocity-sensitive sensory endings in muscle spindles (Brown et al., 1967; Matthews and Stein, 1969). Note that this scheme links the perception of kinematic and kinetic variables and predicts vibration-induced illusions of both position and force—a

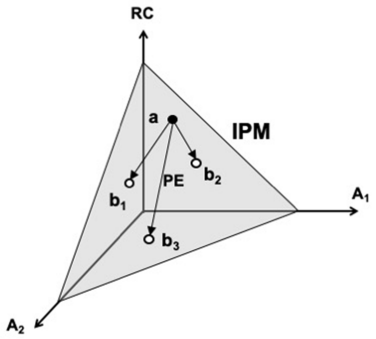


FIGURE 5 | A cartoon illustration of the iso-perceptual manifold (IPM). One coordinate corresponds to an efferent signal (RC), and the other two—to two afferent signals (A_1 and A_2). Note that variations of all signals are possible within the IPM (e.g., from point *a* to points b_1, b_2, and b_3) leading to the undisturbed perception of the salient variable. Such motion is perceptually-equivalent (PE). RC, referent coordinate.

prediction confirmed experimentally (Cafarelli and Kostka, 1981; Reschechtko et al., 2018). Some of the most recent studies explored the potential role of changes in efference copy in kinesthetic illusions, in particular those seen during misperception of force following voluntary muscle coactivation and the drifts in force after turning the visual feedback off (Cuadra et al., 2020, 2021a; Latash, 2021b). Under those conditions, relatively large-amplitude force changes are either not perceived or even perceived as happening in the opposite direction. The authors interpreted those observations as reflections of using distorted efference copy signals. In other words, efference copy is not necessarily a copy of the ongoing efferent process, as suggested earlier based on observations of vibration-induced kinesthetic illusions (Feldman and Latash, 1982).

Some of the mentioned studies also reported differences between two methods used to report percepts: Using verbal reports based on a psychophysical scale and using the contralateral effector to match the perceived variable. Both methods may be seen as suboptimal for obvious reasons such as subjectivity, possible drifts in memorized scales, asymmetry of the effectors on the two sides of the body, and other factors. Those studies observed qualitative differences in the reported percepts based on the two methods (Cuadra et al., 2020, 2021b). For example, coactivating muscles under the instruction to keep the pressing finger force constant leads to an unintentional force increase by about 50%. When asked to report the force change verbally, the subjects report that the force dropped by a small magnitude. In contrast, when asked to match the force with the contralateral hand, the subjects overshoot the actually increased force (Cuadra et al., 2020).

These observations suggest that perceiving-to-report and perceiving-to-act may involve different neural circuits. This conclusion matches well the classical notions of dorsal and ventral brain streams introduced for visual perception (Goodale et al., 1991; Goodale and Milner, 1992; Kravitz et al., 2011). It generalizes these notions to proprioception (see also Proffitt et al., 2003; Zadra et al., 2016) with a possibility that this rule applies to other modalities as well.

ELEMENTS OF PHILOSOPHY OF BIOLOGICAL ACTION

The development of the idea of control with spatial RCs to perception is promising. However, this bottom-up approach may hit serious obstacles when dealing with issues that have traditionally been considered as those of cognition. An attempt to couple cognitive problems, such as, for example, selecting a target for movement, has been made by Gregor Schöner and colleagues in the form of the neural field theory incorporated into a general framework involved in the generation of functional actions, which involves the control with spatial referent coordinates and the synergic control of movements (Erlhagen and Schöner, 2002; Martin et al., 2009, 2019). However, even this most advanced scheme is rather far from dealing with such concepts as *understanding*.

It is possible that another qualitative step is needed to move from the control of biological movements with spatial RCs (and related perceptual phenomena) to issues such as understanding the relations among objects and using this understanding for selection of future motor and non-motor actions, including cognitive actions. This problem seems to be directly related to finding sets of adequate variables for each new level of analysis where variables and methods developed to describe processes at other levels fail (cf. Gelfand, 1991). This problem is also related to the ideas developed by the French philosopher, Merleau-Ponty (1942/1963), of different levels of complexity and associated problems pertaining to processes in inanimate nature ("physical order"), biological systems ("life order"), and conscious systems ("human order").

The theory of control of biological movements with spatial referent coordinates makes a step from laws of nature of the inanimate world to possible laws of nature involved in the motor function of living systems. Can the same basic notions and laws be applied to problems of psychology and cognition? Or, to approach the problem of interface between biological action and cognition from the other side, does the concept of *understanding* apply equally to the fields of animal (including human) movements and to cognitive tasks such as selecting an optimal move in the chess game? Do children *understand* how to use the hand to turn the doorknob when they learn to open the door?

Nikolai Bernstein would probably agree that *understanding* is related to creating a *synergy* within the relevant space of elemental variables although this requires expanding the concept of synergy beyond its definition in his hierarchical scheme for the control of actions (Bernstein, 1947; translation in Latash, 2020b). Within that scheme, Bernstein placed synergy at the second from the bottom level (Level B). Within the same scheme, the concept of understanding (not used in the book) seems to be applicable only at the two top levels, the Level of Actions (Level D) and the Level of Symbolic Actions (Level E). The differences within the hierarchical scheme are one of the factors that make using two words justifiable. So

far, synergy has been linked to action stability but not to selecting targets for action or other decision-making steps. In contrast, the concept of understanding has been developed within a computational approach based on the idea of active inference linked to minimization of variational free energy for a variety of brain functions including the control of movement and decision-making (Friston, 2012; Friston et al., 2013, 2017).

There are several features that are shared by the concepts of synergy and understanding. Both involve organizing the elemental variables into a few basic groups (addressed in movement studies with many names including modes, modules, factors, and primitives, reviewed in Latash, 2020a). Both involve ensuring the stability of task-specific outcomes, which may be picking up a glass with water and moving it to the mouth or finding an optimal move winning the chess game. Indeed, the concept of stability seems highly relevant to understanding: unstable understanding is doubt, which can be equated to the development of or transition toward understanding.

In his most comprehensive book, Bernstein (1947) emphasized the feeling of discovery when learning a skill, which he associated with delegating the responsibility for certain features of the task to lower levels of control (he addressed them as "background levels"), which are typically not perceived consciously. Such discoveries were associated, in particular, with finding dynamically stable trajectories solving the task, i.e., using pre-existing or creating new synergies stabilizing salient variables. For example, after one learns how to ride a bicycle, it is not necessary to think about not falling down, and the brain can become preoccupied with other tasks (e.g., where to ride it to and for what purpose, or even reciting poetry) as long as the road does not present perturbations exceeding the range of dynamical stability.

Using a similar language, *understanding* is also equivalent to delegating certain groups of problems to lower levels such that one is able to take for granted solutions for those problems and to have time and energy to deal with something more exciting and challenging. Can one develop a computational toolbox to measure the ability to understand that could be equivalent to the toolbox associated with the UCM hypothesis described earlier? This would require defining sets of elemental variables, salient higher-level variables, and the mapping rules between the two. A better understanding would imply using broadly varying combinations of elemental variables resulting in acceptable solutions for the cognitive task at hand. Is there an inherent trade-off between understanding (in terms of ensuring the stability of task-solving processes) and optimization (e.g., in terms of energy, Yufik, 2019) similar to the one described earlier for movements (Park et al., 2010)? These are exciting questions without answers so far.

AUTHOR CONTRIBUTIONS

The author is responsible for conceiving the manuscript, analysis, and writing the final draft. The author is responsible for all aspects of work on this article.

ACKNOWLEDGMENTS

The author is grateful to all his former and current students and visiting colleagues who contributed to the studies referred to in this article.

REFERENCES

Akulin, V. M., Carlier, F., Solnik, S., and Latash, M. L. (2019). Sloppy, but acceptable, control of biological movement: algorithm-based stabilization of subspaces in abundant spaces. *J. Hum. Kinet.* 67, 49–72. doi: 10.2478/hukin-2018-0086

Ambike, S., Mattos, D., Zatsiorsky, V. M., and Latash, M. L. (2016a). Synergies in the space of control variables within the equilibrium-point hypothesis. *Neuroscience* 315, 150–161. doi: 10.1016/j.neuroscience.2015.12.012

Ambike, S., Mattos, D., Zatsiorsky, V. M., and Latash, M. L. (2016b). Unsteady steady-states: central causes of unintentional force drift. *Exp. Brain Res.* 234, 3597–3611. doi: 10.1007/s00221-016-4757-7

Ambike, S., Mattos, D., Zatsiorsky, V. M., and Latash, M. L. (2018). Systematic, unintended drifts in the cyclic force produced with the fingertips. *Motor Control* 22, 82–99. doi: 10.1123/mc.2016-0082

Babinski, F. (1899). De l'asynergie cerebelleuse. *Rev. Neurologique* 7, 806–816.

Bernstein, N. A. (1930). A new method of mirror cyclographie and its application towards the study of labor movements during work on a workbench. *Hyg. Safety Pathol. Lab.* 5, 3–9. 6, 3–11.

Bernstein, N. A. (1947). *On the Construction of Movements*. Medgiz: Moscow (in Russian). English translation in Latash, 2020b.

Bernstein, N. A. (1967). *The Co-ordination and Regulation of Movements*. Oxford: Pergamon Press.

Brown, M. C., Engberg, I., and Matthews, P. B. (1967). The relative sensitivity to vibration of muscle receptors of the cat. *J. Physiol.* 192, 773–800. doi: 10.1113/jphysiol.1967.sp008330

Cafarelli, E., and Kostka, C. E. (1981). Effect of vibration on static force sensation in man. *Exp. Neurol.* 74, 331–340. doi: 10.1016/0014-4886(81)90173-4

Cuadra, C., Corey, J., and Latash, M. L. (2021a). Distortions of the efferent copy during force perception: a study of force drifts and effects of muscle vibration. *Neuroscience* 457, 139–154. doi: 10.1016/j.neuroscience.2021.01.006

Cuadra, C., Gilmore, R., and Latash, M. L. (2021b). Finger force matching and verbal reports: testing predictions of the iso-perceptual manifold (IPM) concept. *J. Mot. Behav.* 53, 598–610. doi: 10.1080/00222895.2020.1813681

Cuadra, C., and Latash, M. L. (2019). Exploring the concept of iso-perceptual manifold (IPM): A study of finger force matching. *Neuroscience* 401, 130–141. doi: 10.1016/j.neuroscience.2019.01.016

Cuadra, C., Wojnicz, W., Kozinc, Z., and Latash, M. L. (2020). Perceptual and motor effects of muscle co-activation in a force production task. *Neuroscience* 437, 34–44. doi: 10.1016/j.neuroscience.2020.04.023

Diedrichsen, J., Shadmehr, R., and Ivry, R. B. (2010). The coordination of movement: optimal feedback control and beyond. *Trends Cogn. Sci.* 14, 31–39. doi: 10.1016/j.tics.2009.11.004

Erlhagen, W., and Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychol. Rev.* 109, 545–572. doi: 10.1037/0033-295x.109.3.545

Falaki, A., Huang, X., Lewis, M. M., and Latash, M. L. (2017). Motor equivalence and structure of variance: multi-muscle postural synergies in Parkinson's disease. *Exp. Brain Res.* 235, 2243–2258. doi: 10.1007/s00221-017-4971-y

Feldman, A. G. (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II. controllable parameters of the muscle. *Biophysics* 11, 565–578.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ-model) for motor control. *J. Mot. Behav.* 18, 17–54. doi: 10.1080/00222895.1986.10735369

Feldman, A. G. (2009). New insights into action-perception coupling. *Exp. Brain Res.* 194, 39–58. doi: 10.1007/s00221-008-1667-3

Feldman, A. G. (2015). *Referent Control of Action and Perception: Challenging Conventional Theories in Behavioral Science*. New York, NY: Springer.

Feldman, A. G. (2016). Active sensing without efference copy: referent control of perception. *J. Neurophysiol.* 116, 960–976. doi: 10.1152/jn.00016.2016

Feldman, A. G. (2019). Indirect, referent control of motor actions underlies directional tuning of neurons. *J. Neurophysiol.* 121, 823–841. doi: 10.1152/jn.00575.2018

Feldman, A. G., and Latash, M. L. (1982). Afferent and efferent components of joint position sense: interpretation of kinaesthetic illusions. *Biol. Cybern.* 42, 205–214. doi: 10.1007/BF00340077

Feldman, A. G., and Latash, M. L. (2005). Testing hypotheses and the advancement of science: recent attempts to falsify the equilibrium-point hypothesis. *Exp. Brain Res.* 161, 91–103. doi: 10.1007/s00221-004-2049-0

Feldman, A. G., and Orlovsky, G. N. (1972). The influence of different descending systems on the tonic stretch reflex in the cat. *Exp. Neurol.* 37, 481–494. doi: 10.1016/0014-4886(72)90091-x

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Mattout, J., and Kilner, J. (2013). Action understanding and active inference. *Biol. Cybern.* 104, 137–160. doi: 10.1007/s00422-011-0424-z

Gelfand, I. M. (1991). Two archetypes in the psychology of man. *Nonlinear Sci. Today* 1, 11–16.

Gelfand, I. M., and Latash, M. L. (1998). On the problem of adequate language in movement science. *Motor Control* 2, 306–313. doi: 10.1123/mcj.2.4.306

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature* 349, 154–156. doi: 10.1038/349154a0

Goodwin, G. M., McCloskey, D. I., and Matthews, P. B. (1972). The contribution of muscle afferents to kinaesthesia shown by vibration induced illusions of movement and by the effects of paralysing joint afferents. *Brain* 95, 705–748. doi: 10.1093/brain/95.4.705

Gribble, P. L., and Ostry, D. J. (2000). Compensation for loads during arm movements using equilibrium-point control. *Exp. Brain Res.* 135, 474–482.

Henneman, E., Somjen, G., and Carpenter, D. O. (1965). Excitability and inhibitibility of motoneurones of different sizes. *J. Neurophysiol.* 28, 599–620. doi: 10.1152/jn.1965.28.3.599

Hoffer, J. A., and Andreassen, S. (1981). Regulation of soleus muscle stiffness in premammillary cats: intrinsic and reflex components. *J. Neurophysiol.* 45, 267–285. doi: 10.1152/jn.1981.45.2.267

Horak, F. B., Nutt, J. G., and Nashner, L. M. (1992). Postural inflexibility in parkinsonian subjects. *J. Neurol. Sci.* 111, 46–58. doi: 10.1016/0022-510x(92)90111-w

Jeneson, J. A., Taylor, J. S., Vigneron, D. B., Willard, T. S., Carvajal, L., Nelson, S. J., et al. (1990). 1H MR imaging of anatomical compartments within the finger flexor muscles of the human forearm. *Magn. Reson. Med.* 15, 491–496. doi: 10.1002/mrm.1910150316

Jo, H. J., Lucassen, E., Huang, X., and Latash, M. L. (2017). Changes in multi-digit synergies and their feed-forward adjustments in multiple sclerosis. *J. Mot. Behav.* 49, 218–228. doi: 10.1080/00222895.2016.1169986

Jobin, A., and Levin, M. F. (2000). Regulation of stretch reflex threshold in elbow flexors in children with cerebral palsy: a new measure of spasticity. *Dev. Med. Child Neurol.* 42, 531–540. doi: 10.1017/s0012162200001018

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* 9, 718–727. doi: 10.1016/s0959-4388(99)00028-8

Kravitz, D. J., Saleem, K. S., Baker, C. I., and Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* 12, 217–230. doi: 10.1038/nrn3008

Lackner, J. R., and Taublieb, A. B. (1984). Influence of vision on vibration-induced illusions of limb movement. *Exp. Neurol.* 85, 97–106. doi: 10.1016/0014-4886(84)90164-x

Latash, M. L. (1992). Virtual trajectories, joint stiffness and changes in natural frequency during single-joint oscillatory movements. *Neuroscience* 49, 209–220. doi: 10.1016/0306-4522(92)90089-k

Latash, M. L. (2008). *Synergy*. New York, NY: Oxford University Press.

Latash, M. L. (2010). Motor synergies and the equilibrium-point hypothesis. *Motor Control* 14, 294–322. doi: 10.1123/mcj.14.3.294

Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Exp. Brain Res.* 217, 1–5. doi: 10.1007/s00221-012-3000-4

Latash, M. L. (2018a). Stability of kinesthetic perception in efferent-afferent spaces: the concept of iso-perceptual manifold. *Neuroscience* 372, 97–113. doi: 10.1016/j.neuroscience.2017.12.018

Latash, M. L. (2018b). Muscle co-activation: Definitions, mechanisms and functions. *J. Neurophysiol.* 120, 88–104. doi: 10.2165/00007256-200636020-00004

Latash, M. L. (2019). *Physics of Biological Action and Perception*. New York, NY: Academic Press.

Latash, M. L. (2020a). On primitives in motor control. *Motor Control* 24, 318–346. doi: 10.1123/mc.2019-0099

Latash, M. L. (2020b). *Bernstein's Construction of Movements*. Abingdon, UK: Routledge.

Latash, M. L. (2021a). Laws of nature that define biological action and perception. *Phys. Life Rev.* 36, 47–67. doi: 10.1016/j.plrev.2020.07.007

Latash, M. L. (2021b). Efference copy in kinesthetic perception: a copy of what is it. *J. Neurophysiol.* 125, 1079–1094. doi: 10.1152/jn.00545.2020

Latash, M. L., and Gottlieb, G. L. (1990). Compliant characteristics of single joints: preservation of equifinality with phasic reactions. *Biol. Cybern.* 62, 331–336. doi: 10.1007/BF00201447

Latash, M. L., and Huang, X. (2015). Neural control of movement stability: lessons from studies of neurological patients. *Neuroscience* 301, 39–48. doi: 10.1016/j.neuroscience.2015.05.075

Latash, M. L., Shim, J. K., Smilga, A. V., and Zatsiorsky, V. (2005). A central back-coupling hypothesis on the organization of motor synergies: a physical metaphor and a neural model. *Biol. Cybern.* 92, 186–191. doi: 10.1007/s00422-005-0548-0

Loeb, G. E. (2012). Optimal isn't good enough. *Biol. Cybern.* 106, 757–765. doi: 10.1007/s00422-012-0514-6

Madarshahian, S., Letizi, J., and Latash, M. L. (2021). Synergic control of a single muscle: the example of flexor digitorum superficialis. *J. Physiol.* 599, 1261–1279. doi: 10.1113/JP280555

Madeleine, P., Voigt, M., and Mathiassen, S. E. (2008). Cycle to cycle variability in biomechanical exposure among butchers performing a standardized cutting task. *Ergonomics* 51, 1078–1095. doi: 10.1080/00140130801958659

Madeleine, P., and Madsen, T. M. T. (2009). Changes in the amount and structure of motor variability during a deboning process are associated with work experience and neck-shoulder discomfort. *Appl. Ergon.* 40, 887–894. doi: 10.1016/j.apergo.2008.12.006

Mariappan, Y. K., Manduca, A., Glaser, K. J., Chen, J., Amrami, K. K., and Ehman, R. L. (2010). Vibration imaging for localization of functional compartments of the extrinsic flexor muscles of the hand. *J. Magn. Reson. Imaging* 31, 1395–1401. doi: 10.1002/jmri.22183

Martin, V., Reimann, H., and Schöner, G. (2019). A process account of the uncontrolled manifold structure of joint space variance in pointing movements. *Biol. Cybern.* 113, 293–307. doi: 10.1007/s00422-019-00794-w

Martin, V., Scholz, J. P., and Schöner, G. (2009). Redundancy, self-motion and motor control. *Neural Comput.* 21, 1371–1414. doi: 10.1162/neco.2008.01-08-698

Matthews, P. B., and Stein, R. B. (1969). The sensitivity of muscle spindle afferents to small sinusoidal changes of length. *J. Physiol.* 200, 723–743. doi: 10.1113/jphysiol.1969.sp008719

Mattos, D., Latash, M. L., Park, E., Kuhl, J., and Scholz, J. P. (2011). Unpredictable elbow joint perturbation during reaching results in multijoint motor equivalence. *J. Neurophysiol.* 106, 1424–1436. doi: 10.1152/jn.00163.2011

Mattos, D., Schöner, G., Zatsiorsky, V. M., and Latash, M. L. (2015). Motor equivalence during accurate multi-finger force production. *Hum. Move Sci.* 233, 487–502. doi: 10.1007/s00221-014-4128-1

Merleau-Ponty, M. (1942/1963). *The Structure of Behavior*. Boston, MA: Beacon Press.

Mullick, A. A., Musampa, N. K., Feldman, A. G., and Levin, M. F. (2013). Stretch reflex spatial threshold measure discriminates between spasticity and rigidity. *Clin. Neurophysiol.* 124, 740–751. doi: 10.1016/j.clinph.2012.10.008

Ostry, D. J., and Feldman, A. G. (2003). A critical evaluation of the force control hypothesis in motor control. *Exp. Brain Res.* 153, 275–288. doi: 10.1007/s00221-003-1624-0

Park, J., Zatsiorsky, V. M., and Latash, M. L. (2010). Optimality vs. variability: an example of multi-finger redundant tasks. *Exp. Brain Res.* 207, 119–132. doi: 10.1007/s00221-010-2440-y

Perlovsky, L (2016). Physics of mind. *Front. Syst. Neurosci.* 10:84. doi: 10.3389/fnsys.2016.00084

Prilutsky, B. I., and Zatsiorsky, V. M. (2002). Optimization-based models of muscle coordination. *Exer. Sport Sci. Rev.* 30, 32–38. doi: 10.1097/00003677-200201000-00007

Proffitt, D. R., Stefanucci, J., Banton, T., and Epstein, W. (2003). The role of effort in perceiving distance. *Psychol. Sci.* 14, 106–112. doi: 10.1111/1467-9280.t01-1-01427

Reschechtko, S., Cuadra, C., and Latash, M. L. (2018). Force illusions and drifts observed during muscle vibration. *J. Neurophysiol.* 119, 326–336. doi: 10.1152/jn.00563.2017

Reschechtko, S., and Latash, M. L. (2017). Stability of hand force production: I. hand level control variables and multi-finger synergies. *J. Neurophysiol.* 118, 3152–3164. doi: 10.1152/jn.00485.2017

Reschechtko, S., and Latash, M. L. (2018). Stability of hand force production: II. Ascending and descending synergies. *J. Neurophysiol.* 120, 1045–1060. doi: 10.1152/jn.00045.2018

Roll, J. P., and Vedel, J. P. (1982). Kinaesthetic role of muscle afferents in man, studied by tendon vibration and microneurography. *Exp. Brain Res.* 47, 177–190. doi: 10.1007/BF00239377

Schmidt, R. A., and McGown, C. (1980). Terminal accuracy of unexpected loaded rapid movements: evidence for a mass-spring mechanism in programming. *J. Mot. Behav.* 12, 149–161. doi: 10.1080/00222895.1980.10735215

Scholz, J. P., and Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Exp. Brain Res.* 126, 289–306. doi: 10.1007/s002210050738

Schöner, G. (1995). Recent developments and problems in human movement science and their conceptual implications. *Ecol. Psychol.* 8, 291–314. doi: 10.1016/j.jep.2021.114779

Seif-Naraghi, A. H., and Winters, J. M. (1990). "Optimized strategies for scaling goal-directed dynamic limb movements," in *Multiple Muscle Systems. Biomechanics and Movement Organization*, ed J. M. Winters and S. L.-Y. Woo New York: Springer-Verlag, 312–334.

Shadmehr, R., and Wise, S. P. (2005). *The Computational Neurobiology of Reaching and Pointing*. Cambridge, MA: MIT Press.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *J. Comp. Physiol. Psychol.* 43, 482–489. doi: 10.1037/h0055479

Todorov, E., and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nat. Neurosci.* 5, 1226–1235. doi: 10.1038/nn963

Turpin, N. A., Levin, M. F., and Feldman, A. G. (2016). Implicit learning and generalization of stretch response modulation in humans. *J. Neurophysiol.* 115, 3186–3194. doi: 10.1152/jn.01143.2015

Turvey, M. T. (1990). Coordination. *Am. Psychol.* 45, 938–953. doi: 10.1037//0003-066x.45.8.938

Von Holst, E., and Mittelstaedt, H. (1950/1973). "The reafference principle," in *The Behavioral Physiology of Animals and Man. The Collected Papers of Erich von Holst*, ed R. Martin Coral Gables, Coral Gables, FL: University of Miami Press. 139–173.

Wolpert, D. M., Miall, R. C., and Kawato, M. (1998). Internal models in the cerebellum. *Trends Cogn. Sci.* 2, 338–347. doi: 10.1016/s1364-6613(98)01221-2

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M. (2019). The understanding capacity and information dynamics in the human brain. *Entropy (Basel)* 21:308. doi: 10.3390/e21030308

Yufik, Y. M., and Friston, K. (2016). Life and understanding: origins of the understanding capacity in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3390/e21030308

Zadra, J. R., Weltman, A. L., and Proffitt, D. R. (2016). Walkable distances are bioenergetically scaled. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 39–51. doi: 10.1037/xhp0000107

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

MINI REVIEW
published: 23 November 2021
doi: 10.3389/fnsys.2021.729707

Application of Electroencephalography-Based Machine Learning in Emotion Recognition: A Review

*Jing Cai, Ruolan Xiao, Wenjie Cui, Shang Zhang and Guangda Liu**

College of Instrumentation and Electrical Engineering, Jilin University, Changchun, China

Emotion recognition has become increasingly prominent in the medical field and human-computer interaction. When people's emotions change under external stimuli, various physiological signals of the human body will fluctuate. Electroencephalography (EEG) is closely related to brain activity, making it possible to judge the subject's emotional changes through EEG signals. Meanwhile, machine learning algorithms, which are good at digging out data features from a statistical perspective and making judgments, have developed by leaps and bounds. Therefore, using machine learning to extract feature vectors related to emotional states from EEG signals and constructing a classifier to separate emotions into discrete states to realize emotion recognition has a broad development prospect. This paper introduces the acquisition, preprocessing, feature extraction, and classification of EEG signals in sequence following the progress of EEG-based machine learning algorithms for emotion recognition. And it may help beginners who will use EEG-based machine learning algorithms for emotion recognition to understand the development status of this field. The journals we selected are all retrieved from the Web of Science retrieval platform. And the publication dates of most of the selected articles are concentrated in 2016–2021.

Keywords: EEG, machine learning, emotion recognition, feature extraction, classification

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Oksana Zayachkivska,
Danylo Halytsky Lviv National Medical University, Ukraine
Wellington Pinheiro dos Santos,
Federal University of Pernambuco, Brazil

****Correspondence:***
Guangda Liu
gdliu@jlu.edu.cn

Received: *23 June 2021*
Accepted: *08 November 2021*
Published: *23 November 2021*

Citation:
Cai J, Xiao R, Cui W, Zhang S and Liu G (2021) Application of Electroencephalography-Based Machine Learning in Emotion Recognition: A Review. Front. Syst. Neurosci. 15:729707. doi: 10.3389/fnsys.2021.729707

INTRODUCTION

Emotions are the changes in people's psychological and physiological states when they face external stimuli such as sounds, images, smells, temperature, and so on. And it plays a vital role in mental and physical health, decision-making, and social communication. To realize emotion recognition, Ekman regarded emotions as six discrete and measurable states related to physiological information, namely happy, sad, anger, fear, surprise, and disgust (Ekman, 1999; Gilda et al., 2018). Subsequent studies on emotion recognition mostly followed this emotion classification basis, but some researchers had added new emotional states, including neutral, arousal, relaxed (Bong et al., 2012; Selvaraj et al., 2013; Walter et al., 2014; Goshvarpour et al., 2017; Minhad et al., 2017; Wei et al., 2018). Some people had also provided a new classification standard for emotions, including relaxation, mental stress, physical load, mental stress combined with physical load (Mikuckas et al., 2014). The setting that emotions are discretized states makes the emotion recognition can be perfectly realized by classification in machine learning. The overall process of machine learning for emotion recognition is as follows: the subjects' facial expressions, speech sounds, body movements (Kessous et al., 2010), electromyography (EMG), respiration (RSP) (Wei, 2013), galvanic skin

response (GSR) (Tarnowski et al., 2018), blood volume pulsation (BVP), skin temperature (SKT) (Gouizi et al., 2011), photoplethysmographic (PPG) (Lee et al., 2019), electrocardiogram (ECG) (Hsu et al., 2020), heart rate (HR) (Wen et al., 2014) and electroencephalography (EEG) will appear corresponding changes when stimulated by external audio, visual, audio-visual and other stimuli. In addition to the above external factors that will affect the changes in emotions, autonomic nervous system (ANS) activity is viewed as a major component of the emotion response (Kreibig, 2010). Ekman (1992) analyzed six basic emotions by recording six ANS parameters. And Levenson (2014) discussed emotions activate different patterns of ANS response for different emotions.

The above-mentioned physiological information can be collected via specific devices, then features related to emotion states can be extracted after preprocessing the collected data, and finally, emotion recognition will be realized by classifying these features. Compared with external body changes such as facial expressions and speech sounds, the internal physiological information such as EMG, SKT, ANS, and EEG can more genuinely reflect the emotional changes of the subject due to its inability to conceal deliberately. And among the many physiological signals, there are a vast number of researches on collecting EEG, which contains relatively affluent information to recognize emotions through machine learning algorithms. Aim to classify physically disabled people and Autism children's emotional expressions, Hassouneh et al. (2020) achieved a maximum emotion recognition rate of 87.25% using the long short-term memory (LSTM) as the classifier to EEG signals. Aim to classify Parkinson's disease (PD) from healthy controls, Yuvaraj et al. (2014) presented a computational framework using emotional information from the brain's electrical activity. Face the situation that the diagnosis of depression almost exclusively depends on doctor-patient communication and scale analysis, which has obvious disadvantages such as patient denial, poor sensitivity, subjective biases, and inaccuracy. Li et al. (2019) committed to automatically and accurately depression recognition using the transformation of EEG features and machine learning methods.

This paper summarizes the development of EEG-based machine learning methods for emotion recognition from four aspects: acquisition, preprocessing, feature extraction, and feature classification. It is helpful for beginners who rely upon EEG-based machine learning algorithms for emotion recognition to understand the current development of the field and then find their breakthrough points in this field.

ACQUISITION OF ELECTROENCEPHALOGRAPHY SIGNALS FOR EMOTION RECOGNITION

There are generally two ways to acquire EEG signals related to emotions. One way is to stimulate the subject to produce emotional changes by playing audio, video, or other materials and obtain the EEG signal through the EEG device worn by the subject. Yuvaraj et al. (2014) obtained EEG data using the Emotive EPOC 14-channel EEG wireless recording headset (Emotive Systems, Inc., San Francisco, CA) with 128 Hz sampling frequency per channel from 20 PD patients and 20 healthy by inducing the six basic emotions of happiness, sadness, fear, anger, surprise, and disgust using multimodal (audio and visual) stimuli. Bhatti et al. (2016) used music tracks as stimuli to evoke different emotions and created a new dataset of EEG signals in response to audio music tracks using the single-channel EEG headset (Neurosky) with a sampling rate 512 Hz. Chai et al. (2016) recorded EEG signals related to audio-visual stimuli using a Biosemi Active Two system. And EEG signals were digitized by a 24-bit analog-digital converter with a 512 Hz sampling rate. Chen et al. (2018) used a 16-lead Emotiv brainwave instrument (14 of which were EEG acquisition channels and 2 of which were reference electrodes) at a frequency of 128 Hz. Later, Seo et al. (2019) used a video stimulus to evoke boredom and non-boredom and collected EEG data using the Muse EEG headband from 28 Korean adult participants. And Li et al. (2019) conducted an experiment based on emotional face stimuli and recorded 28 subjects' EEG data from 128-channel HydroCel Geodesic Sensor Net by Net Station software. In Hou et al. (2020), the Cerebus system (Blackrock Microsystems, United States) was used to collect EEG data at a 1 kHz sampling rate using a 32-channel EEG cap. In the same year, Maeng et al. (2020) introduced a new multimodal dataset via Biopac's M150 equipment called MERTI-Apps based on Asian physiological signals. And Gupta et al. (2020) used an HTC Vive VR display to enable participants to interact with immersive 360° videos in VR and collected EEG signals using a 16-channel OpenBCI EEG Cap with a 125 Hz sampling frequency. Later, Keelawat et al. (2021) acquired EEG data based on a Waveguard EEG cap with a 250 Hz sampling rate from 12 students from Osaka University, to whom song samples were presented. What's more, to effectively collect EEG signals, the attachment position of electrodes for EEG equipment in many studies follows the international 10–20 system (Chai et al., 2016; Seo et al., 2019; Hou et al., 2020; Huang, 2021).

Another way is to use the existing, well-known database in the field of emotion recognition based on EEG, including DEAP (Izquierdo-Reyes et al., 2018), MAHNOB-HCI (Izquierdo-Reyes et al., 2018), GAMEEMO (Özerdem and Polat, 2017), SEED (Lu et al., 2020), LUMED (Cimtay and Ekmekcioglu, 2020), AMIGOS (Galvão et al., 2021), and DREAMER (Galvão et al., 2021). After obtaining the original EEG signal related to emotion states, the following operation is to preprocess the EEG signal to improve the quality of the EEG data.

PREPROCESSING METHOD OF ELECTROENCEPHALOGRAPHY SIGNAL

The raw EEG data collected through EEG equipment is mixed with electronic equipment noise, as well as potential artifacts of electrooculography (EOG), electromyogram (EMG), respiration and body movements. Therefore, a series of preprocessing operations are usually performed before the feature extraction of the EEG signal to improve the signal-to-noise ratio.

Bandpass filters are used by most research institutes as a simple and effective noise removal method. However, since there is no precise regulation on the effective frequency band in the EEG signal, the bandpass filters used in different studies had different cutoff frequencies. Generally, the purpose of setting the low cutoff frequency at about 4 Hz (Özerdem and Polat, 2017; Chao et al., 2018; Pane et al., 2019; Yin et al., 2020) was to remove electrooculography (EOG) artifacts (0–4 Hz) and potential artifacts of respiration and body movements within 0–3 Hz. While some documents set the low cutoff frequency at about 1 Hz (Yuvaraj et al., 2014; Bhatti et al., 2016; Liang et al., 2019; Hou et al., 2020; Keelawat et al., 2021), the purpose of which was to remove the baseline drift (DC component) in the EEG signal and the 1/f noise introduced by the acquire equipment. On the other hand, for high cutoff frequency, most researchers set it to about 45 Hz (Kessous et al., 2010; Yuvaraj et al., 2014; Liang et al., 2019; Yin et al., 2020) to remove the other artifact noises at the high frequencies. While, some recent studies (Hou et al., 2020; Lu et al., 2020; Rahman et al., 2020) set it around 70–75 Hz to preserve more emotion-related features among the EEG to improve the accuracy of emotion recognition.

In addition to using bandpass filters for noise suppression, scholars have also adopted many other excellent methods for preprocessing EEG signals. For example, in the work of Aguiñaga and Ramirez (2018), the Laplacian filter described by Murugappan (2012) was implemented to mitigate the problem that EEG signals were naturally contaminated with noise and artifacts. And then, a blind source separation (BSS) algorithm was implemented to remove redundancy between active elements meanwhile preserve information of non-active elements. And in the study of Chen et al. (2018), the independent component analysis (ICA) was used to suppress noise. Furthermore, the study conducted in Cimtay and Ekmekcioglu (2020) compared three types of smoothing filters (smooth filter, median filter, and Savitzky-Golay) on EEG data and concluded that the most useful filter was the classical Savitzky-Golaly which smoothed the data without distorting the shape of the waves. And the main contribution of Alhalaseh and Alasasfeh (2020) relied on using empirical mode decomposition/intrinsic mode functions (EMD/IMF) and variational mode decomposition (VMD) for signal processing purposes. Besides, Keelawat et al. (2021) used EEGLAB, an open-source MATLAB environment for EEG processing, to remove contaminated artifacts based on ICA.

In addition to removing noise and artifacts, there are other tasks to be done in the preprocessing process. Since the effective frequency band of the EEG signal does not exceed 75 Hz, while the sampling rate of some acquisition devices was even as high as 1,000 Hz, far exceeding the required sampling rate, down-sampling was usually required to reduce the amount of data and increase the execution rate of the algorithm (Chao et al., 2018; Rahman et al., 2020). Besides, to correlate EEG data with brain events easily, the continuously recorded EEG data were usually segmented with time windows of different lengths according to the timestamp of occurrence (Cimtay and Ekmekcioglu, 2020). In addition, considering that the EEG signal is composed of different rhythmic components, including Delta rhythm ($<$ 3 Hz), Theta rhythm (4–7 Hz), Alpha rhythm (8–12 Hz), Beta rhythm (13–30 Hz), and Gamma rhythm ($>$31 Hz), some studies used bandpass filters to separate the rhythm components in the preprocessing stage to facilitate later feature extraction (Yulita et al., 2019).

FEATURE EXTRACTION OF EMOTION-RELATED ELECTROENCEPHALOGRAPHY SIGNALS

Feature extraction is the algorithm of extracting the specific characteristic features from the EEG signals. These distinctive features describe each emotion in a unique way. The complexity of the emotion recognition is also reduced when the complex input signal is converted into a crisp dataset (Hemanth et al., 2018). Ten features from the time domain, frequency domain, and wavelet domain are usually extracted. Features in the frequency domain are including power spectral density (PSD). Features belonging to the time domain include latency to amplitude ratio (LAR), peak-to-peak value, kurtosis, mean value, peak-to-peak time window, and signal power. And features from the wavelet domain are including entropy and energy (Bhatti et al., 2016). Besides, fractal dimension and statistical features were used by Nawaz et al. (2020). And several non-linear features such as correlation dimension (CD), approximate entropy (AP), largest Lyapunov exponent (LLE), higher-order spectra (HOS), and Hurst exponent (HE) had been used widely to characterize the emotional EEG signal (Balli and Palaniappan, 2010; Chua et al., 2011).

To extract features related to emotional states from EEG signals, a large number of researches on feature extraction algorithms have emerged. Chai et al. (2016) proposed a novel feature extraction method called the subspace alignment auto-encoder (SAAE), which combined an auto-encoder network and a subspace alignment solution in a unified framework and took advantage of both non-linear transformation and a consistency constraint. And Özerdem and Polat (2017) used Discrete wavelet transform (DWT) for feature extraction from EEG signals. Later, Li et al. (2018) organized differential entropy features from different channels as two-dimensional maps to train the hierarchical convolutional neural network (HCNN). In the same year, Izquierdo-Reyes et al. (2018) applied the Welch algorithm to estimate the PSD of each EEG channel, using a Hanning window of 128 samples. Soroush et al. (2018) extracted non-linear features from EEG data, and they suggested feature variability through time intervals instead of absolute values of features. What's more, discriminant features were selected using the genetic algorithm (GA). And Chen et al. (2018) leveraged EMD to obtain several intrinsic eigenmode functions and Approximation Entropy (AE) of the first four IMFs as features from EEG signals for learning and recognition. Later, In Chao et al. (2019), the frequency domain, frequency band characteristics, and spatial characteristics of the multichannel EEG signals were combined to construct the multiband feature matrix (MFM). Consider that the rhythmic patterns of an EEG series could differ between subjects and between different mental

states of the same subject, Liang et al. (2019) used a segment-based feature extraction method to obtain EEG features in three domains (frequency, time, and wavelet). In Li et al. (2019), the PSD and activity were extracted as original features using the Auto-regress model and Hjorth algorithm with different time windows. And Qing et al. (2019) used the autoencoder to further process the differential feature to improve the discriminative power of the features. Besides, Yulita et al. (2019) used principal component analysis (PCA) to change most of the original variables that correlate with each other into a set of variables that are smaller and mutually independent. Later Alhalaseh and Alasasfeh (2020) used entropy and Higuchi's fractal dimension (HFD) in the feature extraction stage. And Salankar et al. (2021) first adapted EMD to decomposes the signals into several oscillatory IMF and then extracted features including area, mean, and central tendency measure of the elliptical region from second-order difference plots (SODP). In the same year, Wang et al. (2021) proposed an emotion quantification analysis (EQA) method, which was conducted based on the emotional similarity quantification (ESQ) algorithm in which each emotion was mapped in the valence-arousal domains according to the emotional similarity matrixes.

After feature extraction, some studies also reduced the feature space by feature selection (FS) technique to avoid over-specification using large number of extracted features and to make the feature extraction feasible online. In study of Jirayucharoensak et al. (2014), the input features of the deep learning network (DLN) were power spectral densities of 32-channel EEG signals from 32 subjects. To alleviate the overfitting problem, PCA was applied to extract the most important components of initial input features. Later, Rahman et al. (2020) implemented spatial PCA to reduce signal dimensionality and to select suitable features based on the t-statistical inferences. And Zhang et al. (2020) proposed a shared-subspace feature elimination (SSFE) approach to identify EEG variables with common characteristics across multiple individuals. Yin et al. (2020) proposed a new locally robust feature selection (LRFS) method to determine generalizable features of EEG within several subsets of accessible subjects. Besides, Maeng et al. (2020) used GA to determine the active feature group from the extracted features. Also, other FS algorithms, including correlation ratio (CR), mutual information (MI), and random forest (RF), were used in Suzuki et al. (2021). After extracting the emotional state-related feature vectors from the EEG signal, the next important step is to classify these features to achieve emotion recognition.

CLASSIFICATION OF EMOTION-RELATED ELECTROENCEPHALOGRAPHY SIGNALS

The concept of classification is to construct a classifier based on existing data. The classifier is a general term for the methods of classifying samples, and for emotion recognition using EEG signals, it is a crucial part, which takes the features extracted in the above process as input to complete the recognition of the emotional states.

Many classifiers have been implemented to help emotion recognition, including Support Vector Machine (SVM), multilayer perceptron (MLP), Circular Back Propagation Neural Network (CBPN), Deep Kohonen Neural Network (DKNN), deep belief networks with glia chains (DBN-GCs), artificial neural network (ANN), linear discriminant analysis (LDA), capsule

TABLE 1 | Classifiers and their performance.

Classification Item	Author	Model	Accuracy (%)
Arousal and valence	Jirayucharoensak et al., 2014	DLN	Arousal: 46.03 Valence: 49.52
	Choi and Kim, 2018	LSTM	Arousal: 74.65 Valence: 78.00
	Chao et al., 2018	DBN-GCs	Arousal: 75.92 Valence: 76.83
	Maeng et al., 2020	GA-LSTM	Arousal: 94.8 Valence: 91.3
	Keelawat et al., 2021	CNN	Arousal: 56.85 Valence: 73.34
Arousal, valence, and dominance	Chao et al., 2019	CapsNet	Arousal: 68.285 Valence: 66.73 Dominance: 67.25
Positive and negative	Özerdem and Polat, 2017	MLP	77.14
	Lu et al., 2020	SVM	85.11
Positive, negative, and neutral	Li et al., 2018	HCNN	97
	Rahman et al., 2020	ANN	86.57 ± 4.08
	Wang et al., 2020	CNN	90.59
Boredom and non-boredom	Seo et al., 2019	KNN	86.73
Pleasand and unpleasand	Gupta et al., 2020	KNN, SVM	KNN:96.5 SVM:83.7
Happy, calm, sad, and fear	Chen et al., 2018	DBN-SVM	87.32
Happy, sad, angry, and astounded	Li et al., 2019	SVM	89.02
Happy, angry, sad, and relaxed	Pane et al., 2019	RF	75.6
	Kessous et al., 2010	DKNN, CBPN	95–98
Sad, disgust, angry, and surprise	Sakalle et al., 2021	LSTM	94.12
Happy, fear, sad, and neutral	Galvão et al., 2021	MSFBEL	74.22
Happy, sad, surprise, fear, disgust, and angry	Hassouneh et al., 2020	LSTM	87.25

network (CapsNet), convolutional neural network (CNN), multi-scale frequency bands ensemble learning (MSFBEL) and so on. And their emotion recognition accuracies are listed in **Table 1**.

Liu et al. (2020) by combining the CNN, SAE, and DNN and training them separately, the proposed network is shown as an efficient method with a faster convergence than the conventional CNN. And, for the SEED dataset, the best recognition accuracy reaches 96.77%. Topic and Russo (2021) propose a new model for emotion recognition based on the topographic (TOPO-FM) and holographic (HOLO-FM) representation of EEG signal characteristics. Experimental results show that the proposed methods can improve the emotion recognition rate on the different size datasets.

Unlike researches listed in **Table 1**, which only identified a limited set of emotional states (e.g., happiness, sadness, anger, etc.), Galvão et al. (2021) were dedicated to predicting the exact values of valence and arousal in a subject-independent scenario. The systematic analysis revealed that the best prediction model was a KNN regressor ($K = 1$) with Manhattan distance, features from the alpha, beta, gamma bands, and the differential asymmetry from the alpha band. Results, using the DEAP, AMIGOS, and DREAMER datasets, showed that this model could predict valence and arousal values with a low error (MAE < 0.06, RMSE < 0.16).

CONCLUSION AND DISCUSSION

To improve the accuracy of EEG signal-based machine learning algorithms in emotion recognition, researchers have made a lot of efforts in the acquisition, preprocessing, feature extraction, and classification of EEG signals. From the above summary, it can be found that the current stage of emotion recognition based on machine learning is mainly focused on the improvement of accuracy. What's more, some combinations of feature extraction algorithms and classifiers can even achieve 100% accuracy in the two-classification of emotion recognition. And we believe that the following two goals that need to be achieved in emotion recognition based on machine learning are: (1) Perception of smaller changes in emotion; (2). Reduction in the complexity of emotion recognition algorithms so that the algorithm can be transplanted to wearable devices to realize real-time emotion recognition.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work described in this manuscript was supported by the Science and Technology Development Plan Project of Jilin Province (20190303043SF) and the "13th Five-Year Plan" Science and Technology Project of the Education Department of Jilin Province (JJKH20200964KJ).

REFERENCES

Aguiñaga, A. R., and Ramirez, M. A. L. (2018). Emotional states recognition, implementing a low computational complexity strategy. *Health Informatics J.* 24, 146–170. doi: 10.1177/1460458216661862

Alhalaseh, R., and Alasasfeh, S. (2020). Machine-learning-based emotion recognition system using EEG signals. *Computers* 9:95. doi: 10.3390/computers9040095

Balli, T., and Palaniappan, R. (2010). Classification of biological signals using linear and nonlinear features. *Physiol. Meas.* 31, 903–920. doi: 10.1088/0967-3334/31/7/003

Bhatti, A. M., Majid, M., Anwar, S. M., and Khan, B. (2016). Human emotion recognition and analysis in response to audio music using brain signals. *Comput. Hum. Behav.* 65, 267–275. doi: 10.1016/j.chb.2016.08.029

Bong, S. Z., Murugappan, M., and Yaacob, S. (2012). "Analysis of electrocardiogram (ECG) signals for human emotional stress classification," in *Communications in Computer and Information Science*, eds S. G. Ponnambalam, J. Parkkinen, and K. C. Ramanathan (Berlin: Springer), 198–205. doi: 10.1007/978-3-642-35197-6_22

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chao, H., Dong, L., Liu, Y., and Lu, B. (2019). Emotion Recognition from Multiband EEG Signals Using CapsNet. *Sensors* 19:2212. doi: 10.3390/s19092212

Chao, H., Zhi, H., Dong, L., and Liu, Y. (2018). Recognition of Emotions Using Multichannel EEG Data and DBN-GC-Based Ensemble Deep Learning Framework. *Comput. Intell. Neurosci.* 2018:9750904.

Chen, T., Ju, S., Yuan, X., Elhoseny, M., Ren, F., Fan, M., et al. (2018). Emotion recognition using empirical mode decomposition and approximation entropy. *Comput. Electr. Eng.* 72, 383–392. doi: 10.1016/j.compeleceng.2018.09.022

Choi, E. J., and Kim, D. K. (2018). Arousal and valence classification model based on long short-term memory and DEAP data for mental healthcare management. *Healthc. Inform. Res.* 24, 309–316. doi: 10.4258/hir.2018.24.4.309

Chua, K. C., Chandran, V., Acharya, U. R., and Lim, C. M. (2011). Application of higher order spectra to identify epileptic EEG. *J. Med. Syst.* 35, 1563–1571. doi: 10.1007/s10916-010-9433-z

Cimtay, Y., and Ekmekcioglu, E. (2020). Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition. *Sensors* 20:2034. doi: 10.3390/s20072034

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Ekman, P. (1999). "Basic emotions," in *Handbook of Cognition and Emotion, Vol.1*, eds T. Dalgleish and M. J. Power (Hoboken: John Wiley & Sons Ltd), 45–60.

Galvão, F., Alarcão, S. M., and Fonseca, M. J. (2021). Predicting Exact Valence and Arousal Values from EEG. *Sensors* 21:3414. doi: 10.3390/s21103414

Gilda, S., Zafar, H., Soni, C., and Waghurdekar, K. (2018). "Smart music player integrating facial emotion recognition and music mood recommendation," in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, (Chennai: IEEE), 154–158. doi: 10.1109/WiSPNET.8299738

Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2017). An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomed. J.* 40, 355–368. doi: 10.1016/j.bj.2017.11.001

Gouizi, K., Bereksi Reguig, F., and Maaoui, C. (2011). Emotion recognition from physiological signals. *J. Med. Eng. Technol.* 35, 300–307. doi: 10.3109/03091902.2011.601784

Gupta, K., Lazarevic, J., Pai, Y. S., and Billinghurst, M. (2020). ""Affectively VR: Towards VR Personalized Emotion Recognition," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, (New York: ACM), 1–4. doi: 10.1145/3385956.3422122

Hassouneh, A., Mutawa, A. M., and Murugappan, M. (2020). Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods. *Inform. Med. Unlocked* 20:100372. doi: 10.1016/j.imu.2020.100372

Hemanth, D. J., Anitha, J., and Son, L. H. (2018). Brain signal based human emotion analysis by circular back propagation and Deep Kohonen Neural Networks. *Comput. Electr. Eng.* 68, 170–180. doi: 10.1016/j.compeleceng.2018.04.006

Hou, H. R., Zhang, X. N., and Meng, Q. H. (2020). Odor-induced emotion recognition based on average frequency band division of EEG signals. *J. Neurosci. Methods* 334:108599. doi: 10.1016/j.jneumeth.2020.108599

Hsu, Y. L., Wang, J. S., Chiang, W. C., and Hung, C. H. (2020). Automatic ECG-Based Emotion Recognition in Music Listening. *IEEE Trans. Affect. Comput.* 11, 85–99. doi: 10.1109/TAFFC.2017.2781732

Huang, C. (2021). Recognition of psychological emotion by EEG features. *Netw. Model. Anal. Health Inform. Bioinform.* 10:12. doi: 10.1007/s13721-020-00283-2

Izquierdo-Reyes, J., Ramirez-Mendoza, R. A., Bustamante-Bello, M. R., Pons-Rovira, J. L., and Gonzalez-Vargas, J. E. (2018). Emotion recognition for semi-autonomous vehicles framework. *Int. J. Interact. Des. Manuf.* 12, 1447–1454. doi: 10.1007/s12008-018-0473-9

Jirayucharoensak, S., Pan-Ngum, S., and Israsena, P. (2014). EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. *Sci. World J.* 2014:627892. doi: 10.1155/2014/627892

Keelawat, P., Thammasan, N., Numao, M., and Kijsirikul, B. (2021). A Comparative Study of Window Size and Channel Arrangement on EEG-Emotion Recognition Using Deep CNN. *Sensors* 21:1678. doi: 10.3390/s21051678

Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Interfaces* 3, 33–48. doi: 10.1007/s12193-009-0025-5

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biol. Psychol.* 84, 394–421. doi: 10.1016/j.biopsycho.2010.03.010

Lee, M. S., Lee, Y. K., Pae, D. S., Lim, M. T., Kim, D. W., and Kang, T. K. (2019). Fast emotion recognition based on single pulse PPG signal with convolutional neural network. *Appl. Sci.* 9:3355. doi: 10.3390/app9163355

Levenson, R. W. (2014). The Autonomic Nervous System and Emotion. *Emot. Rev.* 6, 100–112. doi: 10.1177/1754073913512003

Li, J., Zhang, Z., and He, H. (2018). Hierarchical Convolutional Neural Networks for EEG-Based Emotion Recognition. *Cogn. Comput.* 10, 368–380. doi: 10.1007/s12559-017-9533-x

Li, X., Zhang, X., Zhu, J., Mao, W., Sun, S., Wang, Z., et al. (2019). Depression recognition using machine learning methods with different feature generation strategies. *Artif. Intell. Med.* 99:101696. doi: 10.1016/j.artmed.2019.07.004

Liang, Z., Oba, S., and Ishii, S. (2019). An unsupervised EEG decoding system for human emotion recognition. *Neural Netw.* 116, 257–268. doi: 10.1016/j.neunet.2019.04.003

Liu, J. X., Wu, G. P., Luo, Y. L., Qiu, S. H., Yang, S., Li, W., et al. (2020). EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder. *Front. Syst. Neurosci.* 14:43. doi: 10.3389/fnsys.2020.00043

Lu, Y., Wang, M., Wu, W., Han, Y., Zhang, Q., and Chen, S. (2020). Dynamic entropy-based pattern learning to identify emotions from EEG signals across individuals. *Measurement* 150:107003. doi: 10.1016/j.measurement.2019.107003

Maeng, J. H., Kang, D. H., and Kim, D. H. (2020). Deep Learning Method for Selecting Effective Models and Feature Groups in Emotion Recognition Using an Asian Multimodal Database. *Electronics* 9:1988. doi: 10.3390/electronics9121988

Mikuckas, A., Mikuckiene, I., Venckauskas, A., Kazanavicius, E., Lukas, R., and Plauska, I. (2014). Emotion recognition in human computer interaction systems. *Elektron. Elektrotech.* 20, 51–56. doi: 10.5755/j01.eee.20.10.8878

Minhad, K. N., Ali, S. H. M. D., and Reaz, M. B. I. (2017). A design framework for human emotion recognition using electrocardiogram and skin conductance response signals. *J. Eng. Sci. Technol.* 12, 3102–3119.

Nawaz, R., Cheah, K. H., Nisar, H., and Yap, V. V. (2020). Comparison of different feature extraction methods for EEG-based emotion recognition. *Biocybern. Biomed. Eng.* 40, 910–926. doi: 10.1016/j.bbe.2020.04.005

Özerdem, M. S., and Polat, H. (2017). Emotion recognition based on EEG features in movie clips with channel selection. *Brain Inform.* 4, 241–252. doi: 10.1007/s40708-017-0069-3

Pane, E. S., Wibawa, A. D., and Purnomo, M. H. (2019). Improving the accuracy of EEG emotion recognition by combining valence lateralization and ensemble learning with tuning parameters. *Cogn. Process.* 20, 405–417. doi: 10.1007/s10339-019-00924-z

Qing, C., Qiao, R., Xu, X., and Cheng, Y. (2019). Interpretable Emotion Recognition Using EEG Signals. *IEEE Access* 7, 94160–94170. doi: 10.1109/ACCESS.2019.2928691

Rahman, M. A., Hossain, M. F., Hossain, M., and Ahmmed, R. (2020). Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal. *Egypt. Inform. J.* 21, 23–35. doi: 10.1016/j.eij.2019.10.002

Sakalle, A., Tomar, P., Bhardwaj, H., Acharya, D., and Bhardwaj, A. (2021). A LSTM based deep learning network for recognizing emotions using wireless brainwave driven system. *Expert Syst. Appl.* 173:114516. doi: 10.1016/j.eswa.2020.114516

Salankar, N., Mishra, P., and Garg, L. (2021). Emotion recognition from EEG signals using empirical mode decomposition and second-order difference plot. *Biomed. Signal Process. Control* 65:102389. doi: 10.1016/j.bspc.2020.102389

Selvaraj, J., Murugappan, M., Wan, K., and Yaacob, S. (2013). Classification of emotional states from electrocardiogram signals: a nonlinear approach based on hurst. *Biomed. Eng. Online* 12:44. doi: 10.1186/1475-925X-12-44

Seo, J., Laine, T. H., and Sohn, K. A. (2019). Machine learning approaches for boredom classification using EEG. *J. Ambient Intell. Humaniz. Comput.* 10, 3831–3846. doi: 10.1007/s12652-019-01196-3

Soroush, M. Z., Maghooli, K., Setarehdan, S. K., and Nasrabadi, A. M. (2018). A novel method of EEG-based emotion recognition using nonlinear features variability and dempster-shafer theory. *Biomed. Eng. Appl. Basis Commun.* 30:1850026. doi: 10.4015/S1016237218500266

Suzuki, K., Laohakangvalvit, T., Matsubara, R., and Sugaya, M. (2021). Constructing an Emotion Estimation Model Based on EEG/HRV Indexes Using Feature Extraction and Feature Selection Algorithms. *Sensors* 21:2910. doi: 10.3390/s21092910

Tarnowski, P., Kołodziej, M., Majkowski, A., and Rak, R. J. (2018). "Combined analysis of GSR and EEG signals for emotion recognition," in *International Interdisciplinary PhD Workshop (IIPhDW)*, (Poland: IEEE), 137–141. doi: 10.1109/IIPHDW.2018.8388342

Topic, A., and Russo, M. (2021). Emotion recognition based on EEG feature maps through deep learning network. *Eng. Sci. Technol.* 24, 1442–1454. doi: 10.1016/j.jestch.2021.03.012

Walter, S., Gruss, S., Limbrecht-Ecklundt, K., Traue, H. C., Werner, P., Al-Hamadi, A., et al. (2014). Automatic pain quantification using autonomic parameters. *Psychol. Neurosci.* 7, 363–380. doi: 10.3922/j.psns.2014.041

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., et al. (2020). Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia* 146:107506. doi: 10.1016/j.neuropsychologia.2020.107506

Wang, L., Liu, H., Zhou, T., Liang, W., and Shan, M. (2021). Multidimensional Emotion Recognition Based on Semantic Analysis of Biomedical EEG Signal for Knowledge Discovery in Psychological Healthcare. *Appl. Sci.* 11:1338. doi: 10.3390/app11031338

Wei, C. Z. (2013). Stress emotion recognition based on RSP and EMG signals. *Adv. Mater. Res.* 709, 827–831. doi: 10.4028/www.scientific.net/AMR.709.827

Wei, W., Jia, Q., Feng, Y., and Chen, G. (2018). Emotion Recognition Based on Weighted Fusion Strategy of Multichannel Physiological Signals. *Comput. Intell. Neurosci.* 2018:5296523. doi: 10.1155/2018/5296523

Wen, W., Liu, G., Cheng, N., Wei, J., Shangguan, P., and Huang, W. (2014). Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Trans. Affect. Comput.* 5, 126–140. doi: 10.1109/TAFFC.2014.2327617

Yin, Z., Liu, L., Chen, J., Zhao, B., and Wang, Y. (2020). Locally robust EEG feature selection for individual-independent emotion recognition. *Expert Syst. Appl.* 162:113768. doi: 10.1016/j.eswa.2020.113768

Yulita, I. N., Julviar, R. R., Triwahyuni, A., and Widiastuti, T. (2019). Multichannel Electroencephalography-based Emotion Recognition Using Machine Learning. *J. Phys. Conf. Ser.* 1230:012008. doi: 10.1088/1742-6596/1230/1/012008

Yuvaraj, R., Murugappan, M., Mohamed Ibrahim, N., Sundaraj, K., Omar, M. I., Mohamad, K., et al. (2014). Detection of emotions in Parkinson's disease using higher order spectral features from brain's electrical activity. *Biomed. Signal Process. Control* 14, 108–116. doi: 10.1016/j.bspc.2014.07.005

Zhang, W., Yin, Z., Sun, Z., Tian, Y., and Wang, Y. (2020). Selecting transferrable neurophysiological features for inter-individual emotion recognition via a shared-subspace feature elimination approach. *Comput. Biol. Med.* 123:103875. doi: 10.1016/j.compbiomed.2020.103875

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

BRIEF RESEARCH REPORT
published: 08 December 2021
doi: 10.3389/fnsys.2021.784404

Evolutionary Advantages of Stimulus-Driven EEG Phase Transitions in the Upper Cortical Layers

Robert Kozma[1]*, *Bernard J. Baars*[2,3] *and Natalie Geld*[4]

[1] Center for Large-Scale Intelligent Optimization and Networks, Department of Mathematics, University of Memphis, Memphis, TN, United States, [2] Center for the Future Mind, Florida Atlantic University, Boca Raton, FL, United States, [3] Society for MindBrain Sciences, San Diego, CA, United States, [4] MedNeuro, Inc., New York, NY, United States

Spatio-temporal brain activity monitored by EEG recordings in humans and other mammals has identified beta/gamma oscillations (20–80 Hz), which are self-organized into spatio-temporal structures recurring at theta/alpha rates (4–12 Hz). These structures have statistically significant correlations with sensory stimuli and reinforcement contingencies perceived by the subject. The repeated collapse of self-organized structures at theta/alpha rates generates laterally propagating phase gradients (phase cones), ignited at some specific location of the cortical sheet. Phase cones have been interpreted as neural signatures of transient perceptual experiences according to the cinematic theory of brain dynamics. The rapid expansion of essentially isotropic phase cones is consistent with the propagation of perceptual broadcasts postulated by Global Workspace Theory (GWT). What is the evolutionary advantage of brains operating with repeatedly collapsing dynamics? This question is answered using thermodynamic concepts. According to neuropercolation theory, waking brains are described as non-equilibrium thermodynamic systems operating at the edge of criticality, undergoing repeated phase transitions. This work analyzes the role of long-range axonal connections and metabolic processes in the regulation of critical brain dynamics. Historically, the near 10 Hz domain has been associated with conscious sensory integration, cortical "ignitions" linked to conscious visual perception, and conscious experiences. We can therefore combine a very large body of experimental evidence and theory, including graph theory, neuropercolation, and GWT. This cortical operating style may optimize a tradeoff between rapid adaptation to novelty vs. stable and widespread self-organization, therefore resulting in significant Darwinian benefits.

Keywords: machine understanding, cortex, perception, consciousness, graph theory, neuropercolation, phase transition, criticality

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Andrew A. Fingelkurts,
BM-Science, Finland
Alessandro E. P. Villa,
University of Lausanne, Switzerland

***Correspondence:**
Robert Kozma
rkozma@memphis.edu

Received: *27 September 2021*
Accepted: *03 November 2021*
Published: *08 December 2021*

Citation:
Kozma R, Baars BJ and Geld N (2021) Evolutionary Advantages of Stimulus-Driven EEG Phase Transitions in the Upper Cortical Layers. Front. Syst. Neurosci. 15:784404. doi: 10.3389/fnsys.2021.784404

1. INTRODUCTION

1.1. Computers, Brains, and Energy

We tend to think of the field of computers and informatics as a major event in the history of ideas, and that is broadly correct. But the mathematics of computation can be traced back to ideas propounded by philosophers and linguists at least a thousand years ago. Western and Asian traditions are often traced to the first millennium BCE; certainly the readable scripts of that time seem to reveal ideas and observations that are remarkably "modern." History is itself a massively parallel distributed network of events over many centuries. It was not until the invention of digital computers about 80 years ago that systematic studies became feasible to explore the possibility of developing man-made intelligent machines (Turing and Haugeland, 1950; Von Neumann, 1958), which have the potential of demonstrating problem-solving performance comparable to humans. Computer technology demonstrated exponential growth for over half a century. Computers support all aspects of our life. Indispensable and pervasive, they lift billions of people out of poverty worldwide and help them to benefit from technological progress in a modern, interconnected society. The dominant approaches in these applications use Neural Networks (NNs) (Barto et al., 1983; Bishop, 1995; Miller et al., 1995) and Deep Learning (DL), and produce cutting-edge AI with often super-human performance (LeCun et al., 2015; Mnih et al., 2015; Schmidhuber, 2015). The present development trend of intelligent technologies is unsustainable. DL has very high demand for computational power and it requires huge data resources, raising many questions from engineering, societal, and ethical perspectives (Jordan and Mitchell, 2015; Marcus, 2018; Kozma et al., 2019a). Computer chips reach hard limits, marked by the approaching end of Moore's law, which dominated computer development for over half a century (Waldrop, 2016). Energy considerations are an important part of the challenges. High-performance computers require increasing proportions of the available electrical energy to operate (Amodei et al., 2018). Moreover, it is increasingly complicated to remove the heat dissipated in the densely packed microchip circuitries.

Brains provide us valuable clues regarding efficient use of resources, including energy. The operation of brains is naturally constrained by the available metabolic resources following fundamental laws of thermodynamics. According to the free energy principle, brains optimize metabolic and computational efficiency by reconfiguring themselves while they interact with the environment in the action and perception cycle (Friston et al., 2006; Sengupta et al., 2013). Brains continuously optimize their energy resource allocation, while advanced computing algorithms are mostly agnostic when it comes to power consumption. Arguably, brains are several orders of magnitude more energy-efficient than cutting-edge AI when solving specific machine learning tasks (Amodei et al., 2018; Kozma et al., 2019b; Marković et al., 2020). The efforts to achieve human-level intelligence and machine understanding by scaling up computing using million-core chips are impressive, but alternative approaches may become useful as well. Energy-awareness is a basic manifestation of embodiment, which is crucial for the emergence of intelligence in brains and machines (Dreyfus, 2007), and it provides the key for progress in machine understanding as well (Yufik, 2013, 2019). Neuromorphic technologies have great potential in large-scale computing systems, including spiking neural networks (Furber, 2016; Hazan et al., 2018; Roy et al., 2019), and memristive hardware (Di Ventra et al., 2009; Chua, 2012; Kozma et al., 2012; Stieg et al., 2019). Combining neuromorphic technologies with brain-inspired thermodynamic models of computing has the potential of providing the required breakthrough in machine understanding (Yufik and Friston, 2016; Friston et al., 2020).

1.2. Cognitive Dynamics and Consciousness

It is often thought that the question of consciousness in the waking brain is so difficult and poorly understood that empirical science has nothing to say about it. However, beginning some decades ago, empirical scientists in psychology and neuroscience have published literally thousands of scientific papers, mostly on very specific aspects of conscious perception and cognition.[1] Global Workspace Theory (GWT) is one of the prominent modeling approaches (Baars, 1997; Baars and Geld, 2019; Baars et al., 2021). GWT fundamentally proposes that the striking capacity limits of conscious percepts implies very widespread unconscious access to processing resources in the brain. This convergence of two very different theoretical traditions suggests that they are two sides of the same coin.

GWT first emerged around 1980, based on the cognitive architecture tradition in cognitive science, including global workspace architecture (Newell et al., 1972). The cognitive architecture program goes back many decades, when Herbert A. Simon and the Netherlands chess master Adrian De Groot began to carefully study the move-by-move "consciousness reports" of advanced chess players (Simon, 1967; De Groot, 2014). Since the middle of the last century, a number of cognitive architectures have been proposed and partially tested. The book by Newell (1994) can be considered to be a summary of this empirical modeling tradition. At least a dozen cognitive architectures have been proposed in this research practice. They proposed different computer implementations with two shared features: All cognitive architectures had a serial perception and problem-solving component, and in all cases the serial flow of immediately accessible events interacted with a very large long-term memory capacity, which appears to be a non-serial set of knowledge sources. Cognitive architectures also merged with a separate experimental cognitive research tradition, until, by the 1970s and 80s, it began to seem that both lines of research could be understood in a single framework (John and Newell, 1990). The work of Tversky and Kahneman (2011) is another example of this pattern of discoveries, focusing on the empirical phenomenon of automaticity. Newell (1994) discussed this striking convergence of a serial "stream of consciousness" reported by subjects, and a very large, non-serial set of memory domains, which are not

[1] In the scientific literature, over 27 thousand relevant abstracts can be found at this link: https://pubmed.ncbi.nlm.nih.gov/?term=conscious+brain.

in reportable consciousness at any given time; but the massively parallel memory domain is unconscious most of the time during chess playing.

Baars was one of the first cognitive scientists to explicitly use the word "conscious" for the serial component of chess-playing protocols, and "unconscious" for the large set of knowledge sources that players demonstrably use, but which may not become explicit in any single chess move. What Baars called Global Workspace Theory (GWT) in the 1980s combined two streams of scientific study, the cognitive architecture tradition and the field of cognitive psychology (Baars, 1997). That convergence seemed to be surprisingly easy to describe. By 1980 the field of cognitive science began to emerge, and the computational, mathematical, and cognitive-behavioral streams of development turned into a single, extensive field of study. Baars' GWT linked a vast empirical literature to the theoretical concept of consciousness, which could be inferred from the mass of evidence, and which also seemed to reflect the reported experiences of subjects in many tasks.

The distinctive feature of all cognitive architectures, including GWT, can be found in Newell's pioneering formulation. Rather than a passive unconscious long-term memory, with more powerful computers the idea emerged that the parallel component reflects a "society" of specialized knowledge sources that were not conscious by themselves, but which interacted to "post messages" on some shared knowledge domain, called a global workspace. Since that time, computational GWT has seen very widespread use in cognitive and computer science. The mathematics of parallel-interactive computation led to both fundamental and practical insights into human cognition. What seemed puzzling and scattered before 1980 gradually emerged with a greater degree of clarity (Franklin et al., 2012).[2]

Cognitive Science is now Cognitive Neuroscience, leading to another large set of converging ideas, with more and more brain and behavioral evidence interacting in fruitful ways. In fields like language studies, for example, it became routine to consider the perceptual aspects of a stream of words (like this one) as conscious, in fast-cycling interaction with multiple unconscious knowledge domains. "Society models" gradually merged with the brain sciences, giving rise to contemporary cognitive neuroscience theory. We prefer to think of a "family" of GWT architectures, where Baars' version is perhaps the best known today, but the family has many members that continue to evolve. Essentially empirical, this set of theories may be considered similar enough to be treated as a "family" of global workspace-like approaches, including (Dehaene et al., 1998; Fingelkurts et al., 2010; Edelman et al., 2011; Tononi and Koch, 2015; Kozma and Freeman, 2016; Mashour et al., 2020; Deco et al., 2021). Each approach is distinctive and each is based on a strong body of evidence; but they converge well. Much to our surprise, a very large scientific literature in neurobiology has also converged with all the fields in a remarkable way.

The current paper presents yet another region of convergence between multiple empirical and theoretical streams of development. With direct brain recordings of the electromagnetic activity of single neurons and massive neuronal networks, we may be seeing a convergence between many intellectual traditions. We view brains as large-scale complex networks, and brain dynamics as percolation processes evolving over these networks, with potentially adaptive structures. We introduce several key analysis methods, such as the thermodynamics of wave packets, statistical physics of criticality and phase transitions, cinematic theory of neurodynamics and metastability, and a hypothesis concerning the interpretation of the experimentally observed neurodynamics using the GWT framework. Two main computational results are introduced to illustrate the findings. The first describes the essential role of non-local axonal connections in maintaining a near-critical state of brain oscillations. The second result concerns the role of astrocyte-neural coupling in maintaining neural fields with rapid transitions between states with high and low synchrony, respectively. We conclude the work with discussing the potential implications of these results to lay down the principles of machine understanding.

The rest of the essay addresses the fundamental question: What could be the evolutionary advantage of brains utilizing phase transitions, as compared to possible alternatives with smooth dynamics?

2. METHODS

Describing brains as open thermodynamic systems converting noisy sensory inputs and metabolic energy into conscious sensory percepts to explicit understanding of the world.

2.1. Thermodynamics of Wave Packets[3]

There is a vast literature on experimental investigations of thermodynamics of brains, see, e.g., Abeles and Gerstein (1988), Fuchs et al. (1992), Freeman (2000), and Friston et al. (2006). Freeman K sets provide a theoretical framework for brain models with a hierarchy of increasingly complex structure, dynamics, and function (Freeman, 1975, 1991, 2000; Kozma and Freeman, 2009). Several key aspects are summarized here, using the concept of metastability,[4] as described in Kozma and Freeman (2016, 2017).

PROPOSITION 1 (Characterization of wave packets (WPs); Kozma and Freeman, 2016). *The action-perception cycle is manifested through the self-organized sequence of metastable,*

[2]Stan Franklin's research group really pushed the world of computer science and AI toward these cognitive architectures and moved the needle into this direction, in a 20+ year strong research program at the University of Memphis.

[3]We take no position on philosophical questions that are often raised in connection with conscious perception, the brain, and the relevance of quantum mechanics and quantum fields. Global workspace theory and neuropercolation should be considered on their respective merits. Both theories have been fruitful, and here we consider how they may interact in interpreting experimental results.

[4]A state of a dynamical system is called metastable, if it is not stable, but it maintains its integrity for an extended period of time, which is meaningful for the analyzed problem. In other words, a metastable state is unstable over very long time scales, but it can be considered stable for shorter, still extended time periods. It is of special interest to study metastability in spatially extended systems, when metastability in time is manifested in the emergence of well-defined spatial patterns for some time periods. Transient dynamics from one metastable state to another metastable state has been extensively studied in various mathematical and physical systems. In the present essay, we refer to metastability appearing in the form of intermittent synchronization of cortical activity.

highly synchronized patterns of spatio-temporal amplitude modulated (AM) activity at the beta/gamma carrier frequency (20-80 Hz). These AM patterns emerge and collapse, and as such they form spatio-temporal Wave Packets (WPs). The WPs evolve as follows:

(i) WPs exist for a time window of ∼100 ms, corresponding to approx. 10Hz frequency band. They have spatially-localized evolving patterns, therefore they are sometimes called wave packets.
(ii) WPs have statistically significant correlations with sensory stimuli and reinforcement contingencies perceived by the subject.
(iii) WPs are separated in time by brief transitionary periods (10-20ms). During these transitionary periods, the AM patterns collapse and large-scale synchrony diminishes.
(iv) The repeated collapse of WPs points to recurring singularities in mammalian cortical dynamics ignited at a given location of the cortex. Following the selection and activation of a Hebbian cell assembly corresponding to the stimulus, the synchronized activity of neural populations rapidly propagates across the cortical sheet in the form of a phase cone.
(v) The rapid transitions and propagation of phase cones following their ignition cannot be explained by synaptic transmissions only, and it requires the emergence of collective dynamics.

The repeated collapse and emergence of the metastable wave packets defines a quasi-periodic oscillatory energy cycle with the following steps:

PROPOSITION 2 (Energy cycle of wave packets; Kozma and Freeman, 2017). *The temporal evolution of Wave Packets is sustained by the corresponding energy cycle, described by thermodynamic processes involving energy and entropy transfer between highly-ordered (liquid) states and disordered (gaseous) states:*

(i) The cycle starts with a disordered background state with low amplitude waves. This state has high entropy and in the thermodynamic sense it is analogous to a gaseous state.
(ii) At a certain space-time point, synchrony is ignited in the neural populations in response to a meaningful stimulus and a phase cone starts to grow from an incipient state. The phase cone develops into a highly structured, metastable WP with low entropy oscillating at a narrow beta/gamma frequency band. The emergence of the WP leads to the dissipation of energy in the form of heat, which is removed by the blood stream through the capillaries. This can be viewed as a condensation process to a liquid state.
(iii) The metastable WP continuously erodes with decreasing synchrony between the neuron components, due to the impact of input stimuli and random perturbations. The entropy increases, which corresponds to the thermodynamic process of evaporation.
(iv) At the end of the cycle, the intensity of the neural firing activity drops to a level when the activity patterns are dissolved and the thermodynamics returns to the high-entropy gaseous state.

This section summarized key aspects of experimental findings on EEG recordings in terms of thermodynamic processes. The next sections introduce methods of statistical physics and mathematical theory of graphs and networks to quantitatively characterize these findings.

2.2. Criticality in Brains and Neuropercolation Model

The thermodynamic interpretation of the action-perception cycle outlined above implies that brains operate through repeated transitions between highly-organized, synchronous states and disorganized states with low levels of synchrony. These observations lead to the hypothesis that brains are critical or near-critical systems, which has been proposed by various authors. One prominent approach is based on the concept of self-organized-criticality (SOC) when a high-dimensional complex system organizes itself to a critical point which is an attractor state. SOC demonstrates scale invariance, including power-law behavior with $1/f$ scaling, where f is the frequency of the events corresponding to the specific problem domains. In the case of neural processes, f could relate, for example, to bursts of spontaneous activity in neural populations, and $1/f$ shows the number of bursts of the given frequency. SOC has been observed in many disciplines, from earthquakes, to solar flares, sandpiles, etc, and in neural tissues as well (Beggs and Timme, 2012; Shew and Plenz, 2013). SOC is widely used now in the interpretation of brain monitoring data, including the connectome, resting state networks, consciousness, and other areas; see, e.g., Fingelkurts et al. (2013), Tagliazucchi (2017), Nosonovsky and Roy (2020), and Wang et al. (2020). Under certain conditions, deviation from the power-law behavior predicted by SOC are observed in brain dynamics, which justify approaches addressing criticality beyond SOC, e.g., critical integration and soft assemblies (Aguilera and Di Paolo, 2021).

A related approach uses percolation theory to describe criticality of brain operation, by modeling the cortical neuropil (Kozma et al., 2005, 2014; Bollobás et al., 2010; Kozma and Puljic, 2015).

PROPOSITION 3 (Neuropercolation model of criticality and phase transition in brain dynamics; Kozma et al., 2005; Kozma and Puljic, 2015). *According to neuropercolation, critical behavior in the cortex is made possible by the filamentous structure of the cortical neuropil, which is the most complex substance in the known universe. Neuropercolation is the generalization of Ising models and lattice cellular automata, and it describes the following aspects of the neuropil:*

(i) Presence of rare long axonal connections between neurons, which allow action at distant locations with minimal delay.
(ii) Contribution of astrocytes cells, which have a key role in metabolic processes and in the formation of field effects.
(iii) Incorporation of random noise effects; the model is robust to noise and noise is an important constructive control parameter to tune the system to achieve desired behavior.
(iv) Input-induced and spontaneous phase transitions between states with large-scale synchrony and without synchrony exhibit brief episodes with long-range spatial correlations.

(v) Neuropercolation proposes a constructive algorithm that self-regulates cortical dynamics at criticality following supercritical explosive excursions.

Beyond the theoretical results, neuropercolation has been employed successfully to interpret experiments with Pavlovian conditioning in rabbits (Kozma et al., 2014; Kozma and Puljic, 2015), on entrainment of sensory processing by respiration in rats and human subjects (Heck et al., 2017, 2019), and strategy changes during learning in gerbils (Kozma et al., 2021).

2.3. Intermittent Metastable Brain Oscillations

There is widespread agreement that processing of sensory information in the cortex is associated with complex spatio-temporal patterns of activity (Abeles, 1982). Experimental observations of intermittent brain oscillations with extended metastable periods, interrupted by rapid transients, are widely discussed in the literature (Lehmann et al., 1987; Buzsáki, 1998). This issue is often framed as a choice between opposing views of continuous vs. discrete cognition. Following the wisdom of Kelso's complementarity principle, the likely answer would be that both discrete and continuous aspects are relevant to cognition through the unity of continuity-discreteness (Fingelkurts and Fingelkurts, 2006; Tognoli and Kelso, 2014; Parr and Friston, 2018). Recent reviews by Josipovic (2019), Menétrey et al. (2021), and Lundqvist and Wutz (2021) help to disentangle the arguments.

The hypothesis that perception happens in discrete epochs has been around for decades, and models of brains as dynamical systems with itinerant trajectories over distributed attractor landscapes provided mathematical tools to support the analysis, see, e.g., Babloyantz and Destexhe (1986), Skarda and Freeman (1987), Freeman (2000), and Tsuda (2001). Crick and Koch (2003) described discrete frames as snapshots in visual processing, as well as in consciousness; while Tetko and Villa (2001) provided evidence of cognitive relevance of spatio-temporal neural activity patterns. The sample-and-hold hypothesis expands on the sampling idea and it describes the perceptual and motor processing cycle (Edelman and Moyal, 2017). Spatiotemporal sequences of time-position patterns have been observed in the human brain associated with cognitive tasks (Tal and Abeles, 2018). Recent models describing sequential processing of complex patterns of brain activity are developed in, e.g., Cabessa and Villa (2018), Malagarriga et al. (2019).

EEG data evaluated using Hilbert analysis also display sudden transitions of cognitive relevance (Brennan et al., 2011; Frohlich et al., 2015), while operational architectonics provides a powerful framework for transient synchronization of operational modules underlying mental states (Fingelkurts et al., 2010, 2017). Phase transitions over large-scale brain networks have been applied to describe the switches from one frame to another in the cinematic theory of neurodynamics and cognition (Kozma and Freeman, 2016, 2017). Kozunov et al. (2018) evaluates MEG visual processing data and points to the role of phase transitions and critical phenomena to understand how meaning can emerge from sensory data. The identified cycle length varies depending on the experimental conditions; i.e., it is in the theta/bands in the cinematic theory (Freeman, 2000; Kozma and Freeman, 2017); while Pereira et al. (2017) estimate a very long cycle of consciousness (2 s). The work by Werbos and Davis (2016) is unique by identifying a very precise clock cycle of 153 ms, by analyzing Buzsáki lab data (Fujisawa et al., 2015).

There are various open issues regarding discrete effects in neurodynamics and some questions were raised about their significance in cognition and consciousness. For example, Fekete et al. (2018) states that the involved brain networks cannot produce switching behavior at the rates observed in brain imaging experiments. They lay out a valuable work, but they do admit that their reasoning does not hold for strongly non-linear systems as brains are. Their proposed multi-scale computation near criticality is certainly interesting and it has a lot in common with the edge of criticality described as the result of ontogenetic development in neuropercolation in the past two decades (Kozma et al., 2005). White (2018) does not question the existence of sudden changes observed by Freeman et al. (2006), Brennan et al. (2011), and Kozma and Freeman (2016), rather it misses the established proof that these neurodynamic effects are relevant to conscious perception. Clearly, there is a need for extensive further experiments before confirming or rejecting the central hypothesis on the key role of phase transitions in cognition and consciousness. Some recent experiments lend support to the hypothesis on discontinuities in cognition, such as entrainment of multi-sensory perception by the respiratory cycle (Heck et al., 2017); how breathing shapes memory functions (Heck et al., 2019); the role of state transitions in strategy changes during an aversive learning paradigm and the formation of Hebbian cell assemblies by identifying emergent causal cortical networks (Kozma et al., 2021); and clustering of phase cones during interictal periods over the epileptogenetic brain region (Ramon and Holmes, 2020). Statistical markers of phase transitions show potential use in psychotherapy (Sulis, 2021).

PROPOSITION 4 (Transient processing in perception; Kozma and Freeman, 2017). *Phase transitions over large-scale brain networks have been applied to describe the switches from one frame to another in the cinematic theory perception, as follows:*

(i) The intermittent emergence and collapse of AM patterns in EEG data is interpreted as the evidence that perceptual information processing happens in discrete steps, aligned with the prominent AM patterns.

(ii) The cinematic theory of perception uses the concept of the frame and the shutter, which follow each other sequentially. There is no exact threshold separating the two phases from each other, rather they transit to each other following the corresponding energy cycle of WP.

(iii) The frames are defined by the dominant AM patterns which are sustained for an extended period of around 100 ms, with significant variation depending on experimental conditions. The frames are selected according to the reinforced contingencies as perceived by the subject. The frame activity is largely synchronous across large cortical areas during the existence of the frame. However, the frame is not

a frozen pattern, rather it oscillates at the beta/gamma carrier frequencies.

(iv) The shutter is defined by the relatively short periods (approx. 10 ms) when the AM patterns collapsed and the neural activity is disordered, still not completely random and maintains some trace of the previous dynamics.

The Freeman/Kozma approach has been called *cinematic*, because the cortical dynamics self organizes into phase plateaus at roughly every ∼ 100 ms, followed by a collapse of the phase plateau for about 10 ms. During the brief collapse of synchrony, the cortex is prepared to receive novel perturbations, while the self organized phase synchrony is a time of relative stability and internal processing. This style of functioning plausibly optimizes a balance between receptivity to novelty and stability, pointing to potential evolutionary advantage by the rapid, moment-to-moment adaptivity of the conscious cortex.

Brains are dynamic systems, they can never stop, not even during the relatively quiet periods when frames with metastable amplitude patterns are maintained. Being constrained to a quasi-periodic attractor basin during a frame is just the sign of relative silence, before the explosive impact of the phase transition, which destroys the existing structure and gives rise to the emergence of a new pattern in response to the new sensory input and its meaning to the subject (Freeman, 2000). Dynamical modeling of the brain includes both continuity of the movement along its trajectory, as well as rapid changes as the path leads from one metastable state to another (Tognoli et al., 2018). The switches are not rigid and they have their own rich dynamic structure and a hierarchy with possibly scale-free distribution (Mora-Sánchez et al., 2019). These results show that an integrative approach to identify major features of cognitive dynamics and consciousness is very productive, including the unity of discrete and continuous operating modalities in brains.

2.4. Hypothesis on the Link Between EEG Perceptual Transition and GWT

Phase transitions and criticality in cortical layers may have a profound impact on the nature of consciousness. There have been various attempts to integrate phase transitions with GWT, such as the one by Werner (2013), to model the emergence of multi-level collective behaviors in brain dynamics. Tagliazucchi (2017) describes consciousness as the integration of fragmented, highly differentiated entities into a unified message, and they use percolation model to describe the propagation of conscious access through the brain network medium, with phase transitions when a critical threshold is reached. Josipovic (2019) elaborates the concept of non-dual awareness in the framework of GWT. GWT is hereby linked to perceptual phase transitions (Freeman, 1991, 2000; Kozma and Freeman, 2016).

PROPOSITION 5 (Main Hypothesis on EEG phase transitions as indications of conscious experience Kozma and Freeman, 2016; Baars and Geld, 2019). *Phase transitions in the cortex are ignited at a given location of the cortex, according to EEG data. Phase transitions generate laterally propagating phase gradients (phase cones) across the cortical sheet. In the context of GWT, these results are interpreted as follows:*

(i) Phase cones are neural signatures of perceptual broadcasts described by GWT.

(ii) The rapid expansion of phase cones, covering large cortical areas within 10-20 ms, are consistent with the propagation of perceptual broadcast postulated by GWT.

(iii) The recurrence time of the cortical phase transitions is about 100 ms, which is consistent with the ∼ 100 ms window identified in numerous perceptual and behavioral experiments.

The ∼ 100 ms time domain has long been studied in the sensory sciences and proposed as an integration period for conscious cortical information processing (Baars, 1988; Madl et al., 2011; Baars and Geld, 2019). GWT suggests that conscious sensory events are the leading edge of adaptation during waking life. The very fast and highly adaptive role of cortex clearly fits within a Darwinian framework of genetic, epigenetic, and moment-to-moment cortical adaptation (Edelman et al., 2011). Edelman's Neural Darwinism is highly consistent with this approach, and specifies the role of selectionism at multiple time and spatial scales in the brain. Interpreting phase cones as neural manifestations of perceptual broadcasts of GWT is an important step to connect the content of consciousness with the temporal structure of consciousness *per se* (Menétrey et al., 2021). Next, computational results are introduced to illustrate the hypothesis.

3. RESULTS

3.1. Long-Axonal Connections Facilitate Criticality in the Neuropil

Brain networks analysis has been successful to study anatomical, functional, and effective brain connectivity, using tools of graph theory (Iglesias and Villa, 2007, 2010; Stam and Reijneveld, 2007; Steyn-Ross and Steyn-Ross, 2010; Bullmore and Sporns, 2012; Haimovici et al., 2013). Imamoglu et al. (2012) suggest that frontal and visual brain regions are part of a functional network that supports conscious object recognition by changes in functional connectivity. Zanin et al. (2021) point out that neuroscience of brain networks often emphasizes the extraction of neural connectivity represented by strong links and highly-connected nodes, although weak links can in fact be critical in determining the transition between universality classes. Most of the existing network-based toolsets extract information on the interaction of localized units and nodes (Korhonen et al., 2021). Brains are metastable systems, and their optimal functioning depends upon a delicate metastable balance between local specialized processes and their global integration (Fingelkurts and Fingelkurts, 2010), while minute perturbations and topological changes can lead to significant deviations from the normal operational dynamics (Tozzi et al., 2017), with an impact on synchronization effects in these complex non-linear systems (Brama et al., 2015; Xu et al., 2020). Random graphs and cellular automata models have been developed for cortical dynamics to address the challenges (Balister et al., 2006; Kozma and Puljic, 2015; Ajazi et al., 2019; Turkheimer et al., 2019). Percolation models are especially helpful in the interpretation of experimental findings describing the intermittent emergence of

common-mode oscillations in neural cell assemblies (Kozma and Freeman, 2016).

An important theoretical finding describes phase transitions in a graph model of the cortical neuropil with a mix of short and long connections, including long axons (Janson et al., 2019). A random graph $G_{\mathbb{Z}_N^2,p}$ is considered over the square grid of size $(N+1)\times(N+1)$, and p is the probability describing random long edges, see Equation (1). We assume periodic boundary conditions, for simplicity, thus we have a torus with the short notation $\mathbb{Z}_N^2$. The set of vertices of G consists of all the vertices of $\mathbb{Z}_N^2$. There are two types of edges E, short and long, respectively. Short edges are all the edges from the torus $\mathbb{Z}_N^2$; i.e., each node has 4 short edges connecting to its 4 direct neighbors. Additionally, we introduce random long edges as follows: for any pair of vertices that are at distance d apart of each other on the lattice, we assign an edge with probability p that depends on the distance:

$$p_d = \mathbb{P}\left((x,y)\in E(G_{\mathbb{Z}_N^2,p}) \text{ and } dist(x,y)=d\right) = c/[Nd^{\alpha}], \quad (1)$$

Here α is a number, e.g., $\alpha = 1$. An activation process is defined on $G_{\mathbb{Z}_N^2,p}$ as follows: Denote by $A(t)$ the set of all active vertices at time t. We say that a vertex v is active at time t if its potential function $\chi_v(t) = 1$ and inactive if $\chi_v(t) = 0$. Therefore, $A(t) = \{v \in V(G) \mid \chi_v(t) = 1\}$. At the start, $A(0)$ consists of all vertices that are active with probability p_0. Each vertex may change its potential based on the states of its neighbors as follows:

$$\chi_v(t+1) = \mathbb{1}\left(\sum_{u\in N(v)} \chi_u(t) \geq k\right) \quad (2)$$

A vertex can become active if at least k of its neighbors are active. Let ρ_t be a proportion of active nodes at time t, i.e., $\rho_t = A(t)/N^2$ then the evolution of ρ_t can be described in a mean-field approximation, for details, see Janson et al. (2019). A key result has been derived for the existence of phase transition of the activation process over $G_{\mathbb{Z}_N^2,p}$:

PROPOSITION 6 (MAIN THEOREM JKRS219: on phase transitions in the neuropercolation model with short and long connections Janson et al., 2019). *For the activation process $A(t)$ over random graph $G_{\mathbb{Z}_N^2,p_d}$, in the mean-field approximation, there exists a critical probability p_c such that for a fixed p, w.h.p.:*

1. *all vertices will eventually be active if $p > p_c$, while*
2. *all vertices will eventually be inactive for $p < p_c$.*
3. *The value of p_c is given as the function of k and λ through the solution of some transcendental equations.*

The main theorem in Proposition 3.1 rigorously proves the existence of phase transitions in neuropercolation model with long axons; its meaning is illustrated in **Figure 1**, using numerical evaluation of the precise mathematical formula. In **Figure 1**, the x-axis shows λ, which scales linearly with the probability of long axons, while the y-axis is the critical probability when the phase transition happens; k indicates the update rule. It is seen that there is a region for small λ values, where the model behaves essentially as a *local system*. For large λ values, the critical probability diminishes what is expected for a *global system* without local order. There is a transitionary region when the incremental addition of long connections does matter, as it is expected to be the case in the neuropil. Clearly, this model cannot grasp all the complexity of brain networks, and there are many advancements including inhibitory and excitatory effects, multi-layer architectures with delayed reentrant connections. Still, the introduced effect is very robust and it is a unique property of the neuropil with a mix of short and long projections. Brains can benefit from the transitionary region for tuning their behavior between local fragmentation and overall global dominance, using adaptation and learning effects.

3.2. Metabolic Processing in the Neuropil Controls Transitions Between States With High and Low Synchrony Based on Hysteresis Dynamics

Following fundamental studies on the brain energy budget (Raichle and Gusnard, 2002; Magistretti, 2006), there are extensive integrative models on metabolic coupling in the neuron-glia ensemble with capillaries (Cloutier et al., 2009; Belanger et al., 2011; Jolivet et al., 2015), and the role of metabolic constraints on spiking activity (Teixeira and Murray, 2015; Zhu et al., 2018; Qian et al., 2019). The models typically use multi-compartmental neuron models, but some simplified still realistic spiking neuron models are popular as well, e.g., Izhikevich (2003).

To describe the emergence of synchronized collective cortical oscillations driven by metabolic constraints, the capillary astrocyte-neuron model (CAN) is introduced, which couples spiking and metabolic processes (Kozma et al., 2018, 2019b). The simplest CAN model has two metabolic variables: $g(t)$ and $m(t)$, where $g(t)$ describes the available glycogen stored in the astrocyte, $m(t)$ models the available ATP in the neuron's mitochondria. Izhikevich (2003) model is used for the spiking neurons, with variables $u(t)$ and $v(t)$, which are the dimensionless membrane potential and the membrane recovery variable, respectively. The following differential equations describe the rate of change of the variables:

$$\begin{aligned} dv/dt &= \Phi_1(u,v) + I(t) \\ du/dt &= \Phi_2(u,v,b^+(m)) \end{aligned} \quad (3a)$$

$$dg/dt = -\Psi_1(g,m) + \kappa \int_{t-\tau}^{t} v(t')dt'$$

$$dm/dt = -\Psi_2(g,m) + \Psi_1(g,m) \quad (3b)$$

Here $\Phi_1(u,v)$ is membrane potential fitting function; $\Phi_2(u,v,m)$ describes the recovery variable dynamics, modulated by the available ATP via $m(t)$. $\Psi_1(g,m)$ and $\Psi_2(g,m)$ describe the attenuation of $g(t)$ and $m(t)$, respectively. $I(t)$ describes the influence of synaptic currents. The integral term in Equation (3b) describes the cumulative effect of spiking on the glutamate concentration in the synaptic cleft, over time period of τ, and κ is a scaling parameter. Izhikevich's model has a sensitivity parameter b regulating spike production inside term $\Phi_2(u,v,m)$. A nominal value of $b = 0.2$ assures regular spiking (Izhikevich, 2003). To close the feedback loop between the metabolic and neural parts of the model, b is modulated by $m(t)$ as follows:

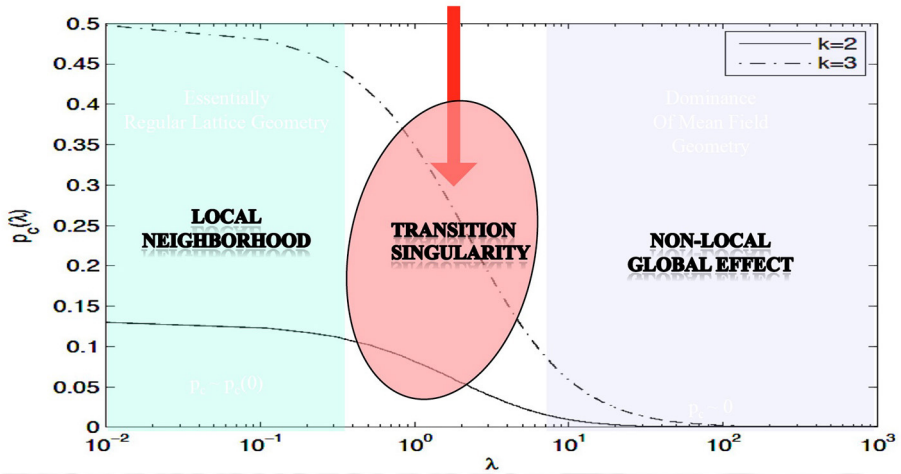


FIGURE 1 | Illustration of the effect of the long edges λ on the critical probability p_c; parameter k specifies the type of the update rule; based on Janson et al. (2019).

$b^+(m(t)) = [\omega b + \beta m(t)]$. Here ω is a scaling parameter in the range [0.75, 1.25], directly impacting the spiking density. The 2nd term reflects the contribution of $m(t)$, where β is a control parameter in the range $[0, 0.5]$. For $\beta = 0$, metabolic processes do not impact spiking, while increasing β leads to increasing frequency of spiking. Full elaboration of the model is given in Kozma et al. (2018).

PROPOSITION 7 (Metabolic control of synchrony transitions in neural populations based on hysteresis dynamics Kozma et al., 2019b). *The capillary astrocyte-neuron model (CAN) described by Equations (3a)–(3b) demonstrates transitions between synchronized collective cortical oscillations and the absence of synchrony, as illustrated in* ***Figure 2****. The process has the following properties:*

(i) The amount of available energy modulates the oscillation frequency of neural populations.
(ii) There is a hysteresis effect as the result of cusp bifurcation in the CAN model. The space defined by the forward gain from neural to metabolic subsystems, and the feedback gain from metabolic to neural system has a bifurcation point leading to the split of a stable equilibrium to two stable and one unstable equilibrium.
(iii) The parameters corresponding to the bifurcated states produce self-sustained oscillations between high and low-synchrony states.
(iv) The results reproduce experimentally observed collective neural dynamics in the form of large-scale cortical phase transitions.

It is important to point out that the metabolic processes are required to produce the hysteresis effect and the desired transitions between states with high and low synchrony. Populations of pure spiking neurons without metabolic components are not sufficient to reproduce the experimentally observed transition effects, as it has been remarked by Deschle et al. (2021).

4. DISCUSSION: HUMAN UNDERSTANDING AND MACHINE UNDERSTANDING

This work explores what the evolutionary advantage may be of brains utilizing repeated phase transitions at theta/alpha rates, as compared to possible alternatives with smooth dynamics. There are a striking number of regularities that are found over and over again at around 10 Hz. Some of these emerge from the mathematics of neurodynamics described here, and some of them emerge from a century of research in conscious sensory perception. We can call this pattern of convergence the "magic number" near-10 Hz (∼100 ms). The flow of conscious events is serial, while unconscious knowledge domains constantly interact with the conscious stream, as EEG data and psychological evidence show over and over again. The ∼100 ms Temporal Window has been studied since the 1800s because it keeps on emerging in psychological evidence. In psychology experiments, it is always linked to highly reliable reports of conscious sensory experiences. As we described here, the magic Temporal Window may be explained by the cinematic view of neurodynamics and phase transitions in the cortex. Because the ∼100 ms Temporal Window is so common, and clearly appears in association with conscious experiences, this possible link is intriguing.

Some of the empirical phenomena that clearly dwell in the magic Temporal Window:

1. Two sensory inputs fuse into single conscious gestalts if they occur within a ∼100 ms time window. This is an enormously general phenomenon in sensory psychophysics, both within and between the major sensory modalities.
2. The motor domain shows a similar Temporal Window. Simple reaction time hovers around ∼100 ms. In continuous tasks, the relationship between sensory output and motor outputs works best within the Temporal Window.
3. The ∼ 100 ms sensory integration window is found in all the major senses, and also in cross sensory tasks. We should

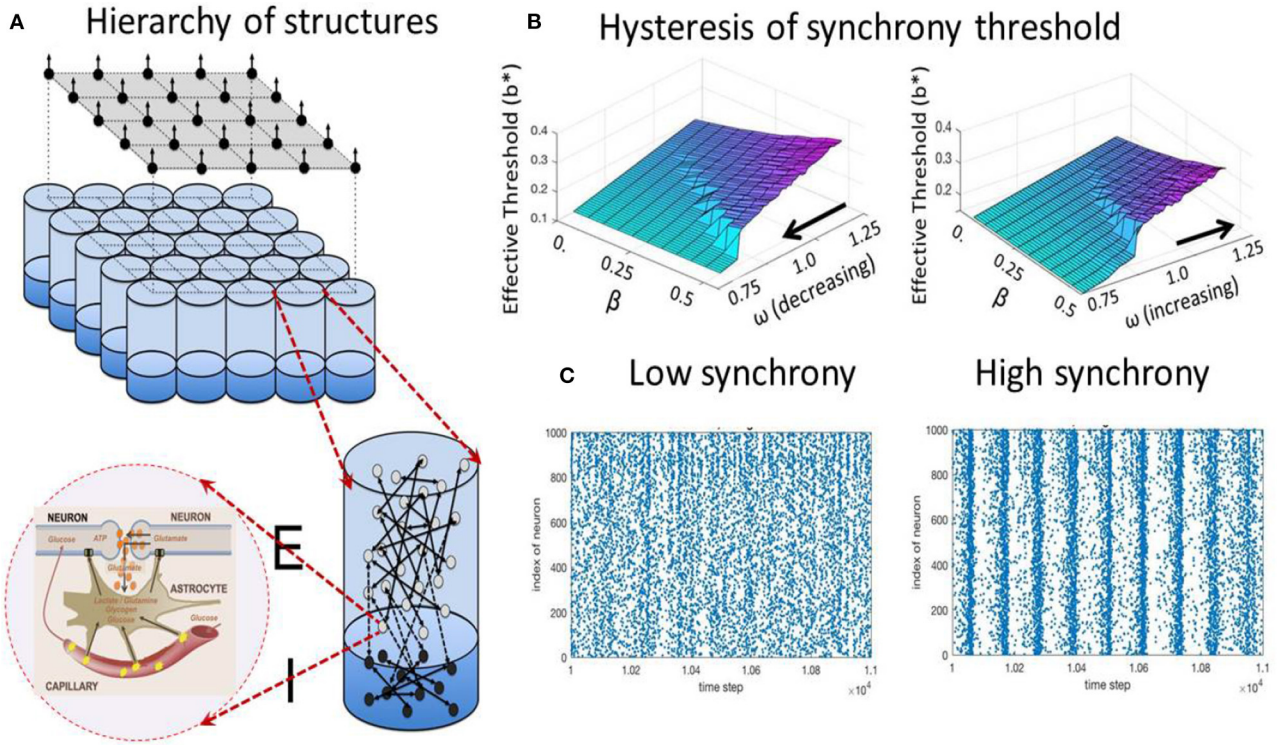


FIGURE 2 | Metabolic-neural model; **(A)** hierarchy of structures from cellular, mini-column, and mesoscopic population levels; **(B)** hysteresis effect on the transition between states with low (blue/turqoise) and high synchrony (purple); **(C)** examples of spiking raster plots for low and high synchrony.

reemphasize the extraordinary generality of this phenomenon across vision, audition, and touch perception in humans and other species. What has been missing is an explanation.

The mathematical properties of cortex, as found by Kozma and Freeman (2016), may therefore explain unconscious-conscious events as they have long been observed in psychology experiments. Phase transitions create the basis for rapid and robust responses to environmental challenges, which provided our ancestors with evolutionary advantage compared to the competitors. As an illustration of these abstract considerations, we can easily imagine a wild rabbit needing to interpret a raptor attack in order to escape it. Under the best possible scenario, it may take $\sim$ 100 ms or more for the rabbit to perceive the attack, and even longer to combine these events with short term and long term memory (Madl et al., 2011). Based on the evolutionary process, this specific time window is sufficient to develop a successful escape strategy while optimizing the finite resources of its brain and body, considering the natural environment, in which the rabbit's ancestors strived for millions of years.

In this work, we outlined a framework for interpreting and modeling brain measurements demonstrating metastable dynamics with rapid transients, which can be used to develop computational devices incorporating brain-inspired principles. Such novel devices have the potential to develop machines which understand the world around us in a way as we humans do, and help us with the challenges we face.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: data are available upon request from the authors. Requests to access these datasets should be directed to rkozma@memphis.edu.

AUTHOR CONTRIBUTIONS

RK coordinated the work with a focus on mathematical and computational modeling and outlined the initial draft manuscript that was modified, and approved by all authors in its final version. BB and NG focused on consciousness and cognitive areas. All authors contributed to the conceptual formulation of this research.

ACKNOWLEDGMENTS

This work was based on the ideas of the paper presented at AFOSR AFRL/RY Workshop on Understanding in the Human and the Machine, 24-26 August 2020, Washington DC. The support of the organizers is greatly appreciated.

REFERENCES

Abeles, M. (1982). *Local Cortical Circuits*. Berlin; Heidelberg: Springer.

Abeles, M., and Gerstein, G. L. (1988). Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *J. Neurophysiol.* 60, 909–924. doi: 10.1152/jn.1988.60.3.909

Aguilera, M., and Di Paolo, E. A. (2021). Critical integration in neural and cognitive systems: beyond power-law scaling as the hallmark of soft-assembly. *Neurosci. Biobehav. Rev.* 123, 230–237. doi: 10.1016/j.neubiorev.2021.01.009

Ajazi, F., Chavez-Demoulin, V., and Turova, T. (2019). Networks of random trees as a model of neuronal connectivity. *J. Math. Biol.* 79, 1639–1663. doi: 10.1007/s00285-019-01406-8

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., and Sutskever, I. (2018). *Ai and Compute*. Heruntergeladen von. Available online at: https://blog.openai.com/aiand-compute

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Stud.* 4, 292–309. doi: 10.1093/acprof:oso/9780195102659.001.1

Baars, B. J., and Geld, N. (2019). *On Consciousness: Science & Subjectivity - Updated Works on Global Workspace Theory*. New York, NY: The Nautilus Press Publishing Group.

Baars, B. J., Geld, N., and Kozma, R. (2021). Global workspace theory (GWT) and prefrontal cortex: Recent developments. *Front. Psychol.* 12:749868. doi: 10.3389/fpsyg.2021.749868

Babloyantz, A., and Destexhe, A. (1986). Low-dimensional chaos in an instance of epilepsy. *Proc. Natl. Acad. Sci. U.S.A.* 83, 3513–3517. doi: 10.1073/pnas.83.10.3513

Balister, P., Bollobás, B., and Kozma, R. (2006). Large deviations for mean field models of probabilistic cellular automata. *Random Struct. Algorithms* 29, 399–415. doi: 10.1002/rsa.20126

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybernet.* 5, 834–846. doi: 10.1109/TSMC.1983.6313077

Beggs, J. M., and Timme, N. (2012). Being critical of criticality in the brain. *Front. Physiol.* 3:163. doi: 10.3389/fphys.2012.00163

Belanger, M. A., and Magistretti, P. (2011). Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell Metab.* 14, 724–738. doi: 10.1016/j.cmet.2011.08.016

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford; New York, NY: Oxford University Press. doi: 10.1201/9781420050646.ptb6

Bollobás, B., Kozma, R., and Miklos, D. (2010). *Handbook of Large-Scale Random Networks*, Vol. 18. Berlin; Heidelberg; New York, NY; Budapest: Springer Science & Business Media.

Brama, H., Guberman, S., Abeles, M., Stern, E., and Kanter, I. (2015). Synchronization among neuronal pools without common inputs: *in vivo* study. *Brain Struct. Funct.* 220, 3721–3731. doi: 10.1007/s00429-014-0886-6

Brennan, B. J., Pockett, S., Bold, G. E. J., and Holmes, M. D. (2011). A possible physiological basis for the discontinuity of consciousness. *Front. Psychol.* 2:377. doi: 10.3389/fpsyg.2011.00377

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336–349. doi: 10.1038/nrn3214

Buzsáki, G. (1998). Memory consolidation during sleep: a neurophysiological perspective. *J. Sleep Res.* 7, 17–23. doi: 10.1046/j.1365-2869.7.s1.3.x

Cabessa, J., and Villa, A. E. (2018). Attractor dynamics of a boolean model of a brain circuit controlled by multiple parameters. *Chaos Interdiscipl. J. Nonlinear Sci.* 28:106318. doi: 10.1063/1.5042312

Chua, L. O. (2012). The fourth element. *Proc. IEEE* 100, 1920–1927. doi: 10.1109/JPROC.2012.2190814

Cloutier, M., Bolger, F., Lowry, J., and Wellstead, P. (2009). An integrative dynamic model of brain energy metabolism using *in vivo* neurochemical measurements. *J. Comp. Neurosci.* 27, 391–414. doi: 10.1007/s10827-009-0152-8

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

De Groot, A. D. (2014). *Thought and Choice in Chess*. The Hague; Paris; New York, NY: De Gruyter Mouton.

Deco, G., Vidaurre, D., and Kringelbach, M. L. (2021). Revisiting the global workspace orchestrating the hierarchical organization of the human brain. *Nat. Hum. Behav.* 5, 497–511. doi: 10.1038/s41562-020-01003-6

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534. doi: 10.1073/pnas.95.24.14529

Deschle, N., Gossn, J. I., Tewarie, P., Schelter, B., and Daffertshofer, A. (2021). On the validity of neural mass models. *Front. Comput. Neurosci.* 14:118. doi: 10.3389/fncom.2020.581040

Di Ventra, M., Pershin, Y. V., and Chua, L. O. (2009). Circuit elements with memory: memristors, memcapacitors, and meminductors. *Proc. IEEE* 97, 1717–1724. doi: 10.1109/JPROC.2009.2021077

Dreyfus, H. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artif. Intell.* 171, 1137–1160. doi: 10.1016/j.artint.2007.10.012

Edelman, G. M., Gally, J. A., and Baars, B. J. (2011). Biology of consciousness. *Front. Psychol.* 2:4. doi: 10.3389/fpsyg.2011.00004

Edelman, S., and Moyal, R. (2017). Fundamental computational constraints on the time course of perception and action. *Prog. Brain Res.* 236, 121–141. doi: 10.1016/bs.pbr.2017.05.006

Fekete, T., Van de Cruys, S., Ekroll, V., and van Leeuwen, C. (2018). In the interest of saving time: a critique of discrete perception. *Neurosci. Conscious.* 4:niy003. doi: 10.1093/nc/niy003

Fingelkurts, A. A., and Fingelkurts, A. A. (2006). Timing in cognition and EEG brain dynamics: discreteness versus continuity. *Cogn. Process.* 7, 135–162. doi: 10.1007/s10339-006-0035-0

Fingelkurts, A. A., and Fingelkurts, A. A. (2010). Alpha rhythm operational architectonics in the continuum of normal and pathological brain states: current state of research. *Int. J. Psychophysiol.* 76, 93–106. doi: 10.1016/j.ijpsycho.2010.02.009

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. (2010). Natural world physical, brain operational, and mind phenomenal space-time. *Phys. Life Rev.* 7, 195–249. doi: 10.1016/j.plrev.2010.04.001

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. (2013). Consciousness as a phenomenon in the operational architectonics of brain organization: criticality and self-organization considerations. *Chaos Solitons Fractals* 55, 13–31. doi: 10.1016/j.chaos.2013.02.007

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. (2017). The legacy of a renaissance man: from mass action in the nervous system and cinematic theory of cognitive dynamics to operational architectonics of brain-mind functioning. *Chaos Complex. Lett.* 11, 81–91.

Franklin, S., Strain, S., Snaider, J., McCall, R., and Faghihi, U. (2012). Global workspace theory, its Lida model and the underlying neuroscience. *Biol. Inspired Cogn. Arch.* 1, 32–43. doi: 10.1016/j.bica.2012.04.001

Freeman, W. (2000). *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*. London: Springer Science & Business Media. doi: 10.1007/978-1-4471-0371-4

Freeman, W. J. (1975). *Mass Action in the Nervous System*. New York, NY: Academic Press.

Freeman, W. J. (1991). The physiology of perception. *Sci. Am.* 264, 78–87. doi: 10.1038/scientificamerican0291-78

Freeman, W. J., Holmes, M. D., West, G. A., and Vanhatalo, S. (2006). Dynamics of human neocortex that optimizes its stability and flexibility. *Int. J. Intell. Syst.* 21, 881–901. doi: 10.1002/int.20167

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., and Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* 118, 42–64. doi: 10.1016/j.neubiorev.2020.07.005

Frohlich, J., Irimia, A., and Jeste, S. S. (2015). Trajectory of frequency stability in typical development. *Brain Imaging Behav.* 9, 5–18. doi: 10.1007/s11682-014-9339-3

Fuchs, A., Kelso, J. S., and Haken, H. (1992). Phase transitions in the human brain: spatial mode dynamics. *Int. J. Bifurcat. Chaos* 2, 917–939. doi: 10.1142/S0218127492000537

Fujisawa, S., Amarasingham, A., Harrison, M. T., Buzsáki, G., and Peyrache, A. (2015). Simultaneous electrophysiological recordings of ensembles of isolated neurons in rat medial prefrontal cortex and intermediate CA1

area of the hippocampus during a working memory task. *CRCNS.org*, 10:K01V5BWK.

Furber, S. (2016). Large-scale neuromorphic computing systems. *J. Neural Eng.* 13:051001. doi: 10.1088/1741-2560/13/5/051001

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110:178101. doi: 10.1103/PhysRevLett.110.178101

Hazan, H., Saunders, D., Khan, J., Sanghavi, D. T., Siegelmann, H. T., and Kozma, R. (2018). BindsNet: a machine learning-oriented spiking neural networks library in Python. *Front. Neuroinform.* 12:89. doi: 10.3389/fninf.2018.00089

Heck, D. H., Kozma, R., and Kay, L. M. (2019). The rhythm of memory: how breathing shapes memory function. *J. Neurophysiol.* 122, 563–571. doi: 10.1152/jn.00200.2019

Heck, D. H., McAfee, S. S., Liu, Y., Babajani-Feremi, A., Rezaie, R., Freeman, W. J., et al. (2017). Breathing as a fundamental rhythm of brain function. *Front. Neural Circ.* 10:115. doi: 10.3389/fncir.2016.00115

Iglesias, J., and Villa, A. E. (2007). Effect of stimulus-driven pruning on the detection of spatiotemporal patterns of activity in large neural networks. *Biosystems* 89, 287–293. doi: 10.1016/j.biosystems.2006.05.020

Iglesias, J., and Villa, A. E. (2010). Recurrent spatiotemporal firing patterns in large spiking neural networks with ontogenetic and epigenetic processes. *J. Physiol. Paris* 104, 137–146. doi: 10.1016/j.jphysparis.2009.11.016

Imamoglu, F., Kahnt, T., Koch, C., and Haynes, J.-D. (2012). Changes in functional connectivity support conscious object recognition. *Neuroimage* 63, 1909–1917. doi: 10.1016/j.neuroimage.2012.07.056

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Netw.* 14, 1569–1572. doi: 10.1109/TNN.2003.820440

Janson, S., Kozma, R., Ruszinkó, M., and Sokolov, Y. (2019). A modified bootstrap percolation on a random graph coupled with a lattice. *Discrete Appl. Math.* 258, 152–165. doi: 10.1016/j.dam.2018.11.006

John, B. E., and Newell, A. (1990). Toward an engineering model of stimulus-response compatibility. *Adv. Psychol.* 65, 427–479. doi: 10.1016/S0166-4115(08)61233-9

Jolivet, R., Coggan, J. S., Allaman, I., and Magistretti, P. J. (2015). Multi-timescale modeling of activity-dependent metabolic coupling in the neuron-glia-vasculature ensemble. *PLoS Comp. Biol.* 11:e1004036. doi: 10.1371/journal.pcbi.1004036

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415

Josipovic, Z. (2019). Nondual awareness: consciousness-as-such as non-representational reflexivity. *Prog. Brain Res.* 244, 273–298. doi: 10.1016/bs.pbr.2018.10.021

Korhonen, O., Zanin, M., and Papo, D. (2021). Principles and open questions in functional brain network reconstruction. *Hum. Brain Mapp.* 42, 3680–3711. doi: 10.1002/hbm.25462

Kozma, R., Alippi, C., Choe, Y., and Morabito, F. C. (2019a). *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. London; Cambridge, MA: Academic Press.

Kozma, R., and Freeman, W. J. (2009). The kiv model of intentional dynamics and decision making. *Neural Netw.* 22, 277–285. doi: 10.1016/j.neunet.2009.03.019

Kozma, R., and Freeman, W. J. (2016). *Cognitive Phase Transitions in the Cerebral Cortex-Enhancing the Neuron Doctrine by Modeling Neural Fields*. Cham: Springer. doi: 10.1007/978-3-319-24406-8

Kozma, R., and Freeman, W. J. (2017). Cinematic operation of the cerebral cortex interpreted via critical transitions in self-organized dynamic systems. *Front. Syst. Neurosci.* 11:10. doi: 10.3389/fnsys.2017.00010

Kozma, R., Hu, S., Sokolov, Y., Wanger, T., Schulz, A. L., Woldeit, M. L., et al. (2021). State transitions during discrimination learning in the gerbil auditory cortex analyzed by network causality metrics. *Front. Syst. Neurosci.* 15:641684. doi: 10.3389/fnsys.2021.641684

Kozma, R., Noack, R., and Manjesh, C. (2018). "Neuroenergetics of brain operation and implications for energy-aware computing," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE), 722–727. doi: 10.1109/SMC.2018.00131

Kozma, R., Noack, R., and Siegelmann, H. T. (2019b). "Models of situated intelligence inspired by the energy management of brains," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (IEEE), 567–572. doi: 10.1109/SMC.2019.8914064

Kozma, R., Pino, R. E., and Pazienza, G. E. (2012). *Advances in Neuromorphic Memristor Science and Applications*, Vol. 4. Dordrecht: Springer Science & Business Media. doi: 10.1007/978-94-007-4491-2

Kozma, R., and Puljic, M. (2015). Random graph theory and neuropercolation for modeling brain oscillations at criticality. *Curr. Opin. Neurobiol.* 31, 181–188. doi: 10.1016/j.conb.2014.11.005

Kozma, R., Puljic, M., Balister, P., Bollobás, B., and Freeman, W. J. (2005). Phase transitions in the neuropercolation model of neural populations with mixed local and non-local interactions. *Biol. Cybernet.* 92, 367–379. doi: 10.1007/s00422-005-0565-z

Kozma, R., Puljic, M., and Freeman, W. J. (2014). "Thermodynamic model of criticality in the cortex based on EEG/ECoG data," in *Criticality in Neural Systems*, eds D. Plenz and E. Niebur (Weinheim: John Wiley & Sons, Ltd.), 153–176. doi: 10.1002/9783527651009.ch7

Kozunov, V., Nikolaeva, A., and Stroganova, T. A. (2018). Categorization for faces and tools-two classes of objects shaped by different experience-differs in processing timing, brain areas involved, and repetition effects. *Front. Hum. Neurosci.* 11:650. doi: 10.3389/fnhum.2017.00650

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lehmann, D., Ozaki, H., and Pál, I. (1987). Eeg alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalogr. Clin. Neurophysiol.* 67, 271–288. doi: 10.1016/0013-4694(87)90025-3

Lundqvist, M., and Wutz, A. (2021). New methods for oscillation analyses push new theories of discrete cognition. *Psychophysiology* e13827. doi: 10.1111/psyp.13827

Madl, T., Baars, B. J., and Franklin, S. (2011). The timing of the cognitive cycle. *PLoS ONE* 6:e14803. doi: 10.1371/journal.pone.0014803

Magistretti, P. J. (2006). Neuron-glia metabolic coupling and plasticity. *J. Exp. Biol.* 209, 2304–2311. doi: 10.1242/jeb.02208

Malagarriga, D., Pons, A. J., and Villa, A. E. (2019). Complex temporal patterns processing by a neural mass model of a cortical column. *Cogn. Neurodyn.* 13, 379–392. doi: 10.1007/s11571-019-09531-2

Marcus, G. (2018). Deep learning: a critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marković, D., Mizrahi, A., Querlioz, D., and Grollier, J. (2020). Physics for neuromorphic computing. *Nat. Rev. Phys.* 2, 499–510. doi: 10.1038/s42254-020-0208-2

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026

Menétrey, M. Q., Vogelsang, L., and Herzog, M. H. (2021). A guideline for linking brain wave findings to the various aspects of discrete perception. *Eur. J. Neurosci.* 2021, 1–10. doi: 10.1111/ejn.15349

Miller, W. T., Werbos, P. J., and Sutton, R. S. (1995). *Neural Networks for Control*. Cambridge, MA: MIT Press.

Mnih, V., Kavukcuoglu, K., Silver, D., and Rusu, A., e. a. (2015). Human-level control through deep reinforcement learning. *Nature* 518:529. doi: 10.1038/nature14236

Mora-Sánchez, A., Dreyfus, G., and Vialatte, F.-B. (2019). Scale-free behaviour and metastable brain-state switching driven by human cognition, an empirical approach. *Cogn. Neurodyn.* 13, 437–452. doi: 10.1007/s11571-019-09533-0

Newell, A. (1994). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., Simon, H. A., et al. (1972). *Human Problem Solving*, Vol. 104. Englewood Cliffs, NJ: Prentice-Hall.

Nosonovsky, M., and Roy, P. (2020). Scaling in colloidal and biological networks. *Entropy* 22:622. doi: 10.3390/e22060622

Parr, T., and Friston, K. J. (2018). The discrete and continuous brain: from decisions to movement and back again. *Neural Comput.* 30, 2319–2347. doi: 10.1162/neco_a_01102

Pereira, A. Jr., Benevides Foz, F., and Freitas da Rocha, A. (2017). The dynamical signature of conscious processing: from modality-specific percepts to complex episodes. *Psychol. Conscious. Theory Res. Pract.* 4:230. doi: 10.1037/cns0000115

Qian, Y., Liu, F., Yang, K., Zhang, G., Yao, C., and Ma, J. (2019). Spatiotemporal dynamics in excitable homogeneous random networks

composed of periodically self-sustained oscillation. *Sci. Rep*. 7, 1–13. doi: 10.1038/s41598-017-12333-3

Raichle, M., and Gusnard, D. (2002). Appraising the brain's energy budget. *Proc. Natl. Acad. Sci. U.S.A*. 99, 10237–10239. doi: 10.1073/pnas.172399499

Ramon, C., and Holmes, M. D. (2020). Increased phase cone turnover in 80-250 Hz bands occurs in the epileptogenic zone during interictal periods. *Front. Hum. Neurosci*. 14:615744. doi: 10.3389/fnhum.2020.615744

Roy, K., Jaiswal, A., and Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 607–617. doi: 10.1038/s41586-019-1677-2

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw*. 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and efficiency in the nervous system-a synthesis. *PLoS Comput. Biol*. 9:e1003157. doi: 10.1371/journal.pcbi.1003157

Shew, W. L., and Plenz, D. (2013). The functional benefits of criticality in the cortex. *Neuroscientist* 19, 88–100. doi: 10.1177/1073858412445487

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychol. Rev*. 74:29. doi: 10.1037/h0024127

Skarda, C. A., and Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behav. Brain Sci*. 10, 161–173. doi: 10.1017/S0140525X00047336

Stam, C. J., and Reijneveld, J. C. (2007). Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomed. Phys*. 1, 1–19. doi: 10.1186/1753-4631-1-3

Steyn-Ross, A., and Steyn-Ross, M. (2010). *Modeling Phase Transitions in the Brain*, Vol. 509. New York, NY; Dordrecht; Heidelberg; London: Springer. doi: 10.1007/978-1-4419-0796-7

Stieg, A. Z., Avizienis, A. V., Sillin, H. O., Aguilera, R., Shieh, H.-H., Martin-Olmos, C., et al. (2019). "Self-organization and emergence of dynamical structures in neuromorphic atomic switch networks," in *Handbook of Memristor Networks*, eds L. Chua, G. Sirakoulis, and A. Adamatzky (Cham: Springer), 391–427. doi: 10.1007/978-3-319-76375-0_14

Sulis, W. (2021). The continuum from temperament to mental illness: dynamical perspectives. *Neuropsychobiology* 80, 135–147. doi: 10.1159/000509572

Tagliazucchi, E. (2017). The signatures of conscious access and its phenomenology are consistent with large-scale brain communication at criticality. *Conscious. Cogn*. 55, 136–147. doi: 10.1016/j.concog.2017.08.008

Tal, I., and Abeles, M. (2018). Imaging the spatiotemporal dynamics of cognitive processes at high temporal resolution. *Neural Comput*. 30, 610–630. doi: 10.1162/neco_a_01054

Teixeira, F. P. P., and Murray, S. (2015). "Local and global criticality within oscillating networks of spiking neurons," in *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 1–7. doi: 10.1109/IJCNN.2015.7280561

Tetko, I. V., and Villa, A. E. (2001). A pattern grouping algorithm for analysis of spatiotemporal patterns in neuronal spike trains. 2. Application to simultaneous single unit recordings. *J. Neurosci. Methods* 105, 15–24. doi: 10.1016/S0165-0270(00)00337-X

Tognoli, E., Dumas, G., and Kelso, J. S. (2018). A roadmap to computational social neuroscience. *Cogn. Neurodyn*. 12, 135–140. doi: 10.1007/s11571-017-9462-0

Tognoli, E., and Kelso, J. S. (2014). The metastable brain. *Neuron* 81, 35–48. doi: 10.1016/j.neuron.2013.12.022

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B Biol. Sci*. 370:20140167. doi: 10.1098/rstb.2014.0167

Tozzi, A., Peters, J. F., Fingelkurts, A. A., Fingelkurts, A. A., and Marijuán, P. C. (2017). Topodynamics of metastable brains. *Phys. Life Rev*. 21, 1–20. doi: 10.1016/j.plrev.2017.03.001

Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behav. Brain Sci*. 24, 793–810. doi: 10.1017/S0140525X01000097

Turing, A. M., and Haugeland, J. (1950). *Computing Machinery and Intelligence*. Cambridge, MA: MIT Press. doi: 10.1093/mind/LIX.236.433

Turkheimer, F. E., Hellyer, P., Kehagia, A. A., Expert, P., Lord, L.-D., Vohryzek, J., et al. (2019). Conflicting emergences. Weak vs. strong emergence for the modelling of brain function. *Neurosci. Biobehav. Rev*. 99, 3–10. doi: 10.1016/j.neubiorev.2019.01.023

Tversky, A., and Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Von Neumann, J. (1958). *The Computer and the Brain*. New Haven, CT; London: Yale University Press.

Waldrop, M. M. (2016). The chips are down for Moore's law. *Nature News* 530:144. doi: 10.1038/530144a

Wang, R., Lin, P., Liu, M., Wu, Y., Zhou, T., and Zhou, C. (2020). Hierarchical connectome modes and critical state jointly maximize human brain functional diversity. *Phys. Rev. Lett*. 123:038301. doi: 10.1103/PhysRevLett.123.038301

Werbos, P. J., and Davis, J. J. (2016). Regular cycles of forward and backward signal propagation in prefrontal cortex and in consciousness. *Front. Syst. Neurosci*. 10:97. doi: 10.3389/fnsys.2016.00097

Werner, G. (2013). Consciousness viewed in the framework of brain phase space dynamics, criticality, and the renormalization group. *Chaos Solitons Fractals* 55, 3–12. doi: 10.1016/j.chaos.2012.03.014

White, P. A. (2018). Is conscious perception a series of discrete temporal frames? *Conscious. Cogn*. 60, 98–12. doi: 10.1016/j.concog.2018.02.012

Xu, Y., Wu, X., Mao, B., Lü, J., and Xie, C. (2020). Fixed-time synchronization in the pth moment for time-varying delay stochastic multilayer networks. *IEEE Trans. Syst. Man Cybernet. Syst*. 52, 1–10. doi: 10.1109/TSMC.2020.3012469

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M. (2019). The understanding capacity and information dynamics in the human brain. *Entropy* 21:308. doi: 10.3390/e21030308

Yufik, Y. M., and Friston, K. (2016). Life and understanding: the origins of "understanding" in self-organizing nervous systems. *Front. Syst. Neurosci*. 10:98. doi: 10.3389/fnsys.2016.00098

Zanin, M., Ivanoska, I., Güntekin, B., Yener, G., Loncar-Turukalo, T., Jakovljevic, N., et al. (2021). A fast transform for brain connectivity difference evaluation. *Neuroinformatics*. 1–15. doi: 10.1007/s12021-021-09518-7

Zhu, Z., Wang, R., and Zhu, F. (2018). The energy coding of a structural neural network based on the Hodgkin-Huxley model. *Front. Neurosci*. 12:122. doi: 10.3389/fnins.2018.00122

Conflict of Interest: NG founded MedNeuro, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 23 December 2021
doi: 10.3389/fnsys.2021.786252

Situational Understanding in the Human and the Machine

Yan Yufik[1]* *and Raj Malhotra*[2]

[1] *Virtual Structures Research, Inc., Potomac, MD, United States,* [2] *United States Air Force Sensor Directorate, Dayton, OH, United States*

The Air Force research programs envision developing AI technologies that will ensure battlespace dominance, by radical increases in the speed of battlespace understanding and decision-making. In the last half century, advances in AI have been concentrated in the area of machine learning. Recent experimental findings and insights in systems neuroscience, the biophysics of cognition, and other disciplines provide converging results that set the stage for technologies of machine understanding and machine-augmented Situational Understanding. This paper will review some of the key ideas and results in the literature, and outline new suggestions. We define situational understanding and the distinctions between understanding and awareness, consider examples of how understanding—or lack of it—manifest in performance, and review hypotheses concerning the underlying neuronal mechanisms. Suggestions for further R&D are motivated by these hypotheses and are centered on the notions of Active Inference and Virtual Associative Networks.

Keywords: understanding, neuronal packet, active inference, complexity, cognitive effort

OPEN ACCESS

Edited by:
Robinson E. Pino,
Office of Science (DOE), United States

Reviewed by:
Maryam Parsa,
George Mason University, United States
Gina Adam,
George Washington University, United States
Todd L. Hylton,
University of California, San Diego, United States

***Correspondence:**
Yan Yufik
imc.yufik@att.net

Received: *30 September 2021*
Accepted: *24 November 2021*
Published: *23 December 2021*

Citation:
Yufik Y and Malhotra R (2021) Situational Understanding in the Human and the Machine. Front. Syst. Neurosci. 15:786252. doi: 10.3389/fnsys.2021.786252

INTRODUCTION: DEFINING SITUATIONAL AWARENESS AND SITUATIONAL UNDERSTANDING

The notions of Situational Awareness and Situational Understanding figure prominently in multiple DoD documents, predicating the achievement of battlespace dominance on SA/SU superiority as, for example, in the following:

> "Joint and Army commanders rely on data, information, and intelligence during operations to develop situational understanding against determined and adaptive enemies. . . because of limitations associated with human cognition, and because much of the information obtained in war is contradictory or false, more information will not equate to better understanding. Commanders and units must be prepared to integrate intelligence and operations to develop situational understanding" (The United States Army Functional Concept for Intelligence, 2020–2040, TRADOC 2017 Pamphlet 525- 2-, p. iii).

Distinctions between SA and SU are defined as follows:

> "Situational awareness is immediate knowledge of the conditions of the operation, constrained geographically in time. More simply, it is Soldiers knowing what is currently happening around them. Situational awareness occurs in Soldiers' minds. It is not a display or the common operating picture; it is the interpretation of displays or the current actual observation of the situation. . . .
>
> Situational understanding is the product of applying analysis and judgment to relevant information to determine the relationships among the mission variables to facilitate decision making. It enables commanders to determine the implications of what is happening and forecast what may happen." The United States Army Operations and Doctrine. Guide to FM-3-0.

Definitive publications by the originator of SA/SU concept and theory (Endsley, 1987, 1988, 1994; Endsley and Connors, 2014) identify three levels of situation awareness and associate understanding with Level 2, as shown in **Figure 1**.

According to the schematic in **Figure 1**, understanding mediates between perception and prediction. The question is: what does such mediation involve, what, exactly, does understanding contribute? The significance of such a contribution can be questioned by, for example, pointing at innumerable cases in the animal domain of going directly from perception to prediction (e.g., intercepting preys requires predators to possess mechanisms for movement prediction, as in frogs shooting their tongues to catch flying insects). The bulk of this paper is dedicated to analyzing the role and contribution of understanding in human performance, pointing, in particular, at uniquely human forms of prediction involving generation of explanations derived from attentively (deliberately, consciously) constructed situation models. Because prediction necessarily entails the consequences of action, these models must include the (counterfactual) consequences of acting. In turn, this mandates generative models of the future (i.e., with temporal depth) and implicit agency. The ensuing approach differs from that adopted in the conventional AI, as follows.

Behaviorist psychology conceptualized the brain as a "black box" and was "fanatically uninterested" in reports concerning events in the box (Solms, 2021, p.10). Borrowing from this expression, one might suggest that cognitivist psychology and AI have been "fanatically uninterested" in the role of understanding; focusing predominantly on learning and reasoning (this contention will be re-visited later in the paper). This paper argues that the capacity for understanding is the definitive feature of human intellect enabling adequate performance in novel situations when one needs to act without the benefit of prior experience or even to counteract the inertia of prior learning. The argument is presented in five parts: the remainder of part I analyzes the notions of situation awareness and situation understanding, focusing on the latter; part II outlines Virtual Associative Network (VAN) theory of understanding, part III places VAN theory in a broader context of Active Inference, part IV considers implementation (machine understanding), followed by a concluding discussion in part V. In the remainder of this part, we define some of the key notions that set the stage for and will be unpacked in the rest of the paper.

The central tenet of this paper boils down to the notion that understanding involves self-directed construction and the manipulation of mental models. In short, planning (as inference). This idea is not original but suggestions concerning the structure of the models and the underlying neuronal mechanisms are (Yufik, 1996, 1998; Yufik and Friston, 2016; Yufik, 2018, 2021b). **Figure 2** introduces some key notions in the proposal, seeking to position mechanisms of awareness and mental modeling within the brain's functional architecture. Stated succinctly, the following treatment builds upon an understanding of the computational architecture of the only systems that evince "understanding"; namely, ourselves.

Figure 2 adopts the classical three-partite model of brain architecture in Luria (1973, 1974); Sigurdsson and Duvarci (2016), except for the inclusion of the cerebellum and Periaqueductal Grey (PAG) structure, whose role in cognitive processes—in particular the maintenance of awareness—was recently discovered (Solms, 2021). It was found that removing the bulk of cortex (in both R and M systems) while leaving the PAG intact preserves a degree of awareness (Solms, 2021). For example, hydranencephalic children (born without cortex) respond to objects placed in their hands, and surgically decorticated animals remain capable of some responses and even rudimentary learning (moreover, in some cases a casual observer might fail to notice differences in the behavior of decorticated animals and intact controls) (Oakley, 1981; Cerminara et al., 2009; Solms, 2021). By contrast, lesions of the PAG and/or reticular structure obliterate awareness (reticular structures project into cortex while PAG receives converging projections from cortex) (Solms, 2021). The architecture in **Figure 2** indicates that intact PAG and RAS support *minimal awareness* (link from PAG to MSP indicates awareness achieved in the absence of the cortex) while an interplay of all the other functional systems produces a hierarchy of awareness levels above the minimal.

Levels of Awareness

To define levels of awareness, one needs to conceptualize the world as generating a stream of stimuli and cognition as a process of assimilating sensory streams aimed to extract energy and sustain energy inflows (these crucially important notions will be re-iterated throughput the paper). With these notions in mind, the following levels of awareness can be identified.

1. *Minimal awareness* (*"vegetative wakefulness,"* the term is due to Solms, 2021, p. 134). Streams of sensory stimuli are experienced as flux (noise).
2. *Selective awareness.* Organism responds to fixed combinations of contiguous stimuli as they appear in the flux (as in frogs catching flies).
3. *Discriminating awareness.* "Blobs" with fuzzy boundaries emerge in perceptual synthesis comprising some contiguous stimuli groupings with varying correlation strength inside the groups.
4. *Differentiating awareness.* Different stimuli compositions are assimilated into "blobs" that are sharply bounded and segregated from the surrounds ("blobs" subsequently turn into distinct "objects," as in telling letters apart).
5. *Recognition-based awareness.* Variations in stimuli compositions in the objects are differentiated (stated differently, different stimuli compositions are experienced as manifestations of the same object, as in recognizing letters in different fonts or handwriting).
6. *Context-based awareness.* The perceptual recognition of objects is influenced by their surrounds (think of the often-cited example of perceiving a shape that looks like a distorted letter A or distorted letter H, depending on its appearance in the middle of C_T or the beginning of _AT).
7. *Understanding-based awareness.* This level is qualitatively different from the preceding levels: all levels deal with learning, i.e., developing memory structures reflecting the

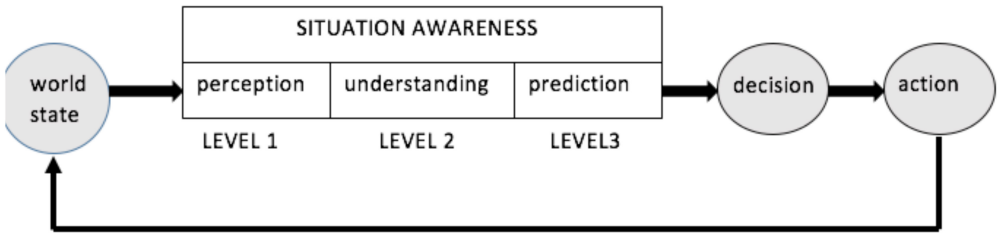


FIGURE 1 | Three levels of Situation Awareness (adopted from Endsley and Connors, 2014).

> statistics of correlation, contingencies and contiguity in the world. By contrast, this level produces and manipulates complex relational structures (mental models) uprooted from such statistics (accounting for non-contiguous and weakly correlated, sparse stimuli groupings) — in other words, compositions and counterfactuals. To illustrate the distinction: the statistics of English texts would suffice for resolving the "_AT or C_T" ambiguity but not for understanding the expression "hats on cats" (when was the last time you saw or read about cats wearing hats?).

Arguably, **Figure 1** refers primarily to understanding-based situation awareness. It is informative to note that cells in prefrontal cortices represent the association of sensory items of more than one sensory modality, integrate these items across time and participate in performing tasks requiring reasoning and manipulation of complex relational structures (Kroger et al., 2002). Construction and manipulation of complex relational structures underlies understanding. More precisely, understanding enables construction of models expressing unlikely correlations (like cats in hats), while sometimes failing to register some precise and routinely encountered ones (e.g., medieval medicine for centuries failed to see the relation between a beating heart and blood circulation, placing the source of circulation in the liver). This paper offers ideas seeking to account for both the strengths and the weaknesses of the understanding capacity. Three pivotal notions (*work, switching, and arousal*) are referenced in **Figure 2** (labeled in italics).

Work (Mental Work)

Operations on mental models demand effort and energy, in the same manner as are those demanded by any bodily (i.e., thermodynamic) work, such as running or lifting weights.

Switching

The functional architecture in **Figure 2** is shared across many species, except for the capability to temporarily decouple mental models from the motor-sensory periphery and environmental feedback. The emergence of this regulatory capacity—to allow such decoupling— underwrites the development of an understanding capacity that is uniquely human and enables uniquely human skills, such as extending the horizon of prediction reach from the immediate to an indefinitely remote future and extending actions reach from objects in the immediate surrounds to indefinitely distant ones, etc.

Arousal

Regulation of arousal (energy distribution in the cortices) is an integral and critical ingredient of mental modeling. In particular, modeling is contingent on maintaining the stability and integrity of neuronal structures in the cortices implementing the models. Resisting entropic erosion and disintegration of the structures require sustained inflows of metabolic energy. These ideas will be given precise definitions that will be mapped onto a simple mathematical formalism.

To summarize, three different brain mechanisms have been identified: those that circumvent the cortex, those that engage the cortex, and those mechanisms that are realized in the cortex and are temporarily disengaged from the motor-sensory periphery (*switching).* The former two mechanisms underlie learning and are shared among multiple species, including humans, while the latter is unique to humans and underlies understanding. The proposal so far is derived from the following conceptualization: (a) the world is a process or stream (not a "static pond"), (b) cognition is a process of adapting an organism's state and behavior to variations in the stream, and (c) the adaptation are powered by energy (*work*) extracted from the stream and distributed inside the system (*regulation of arousal*). Understanding complements learning: learning extrapolates from past experiences, while understanding overcomes the inertia of learning when encountering new conditions with no precedents. Overcoming inertia is an effortful process that can fail but provides unique performance advantages when it succeeds. It was noted that AI and the cognitivist school of thought have downplayed the role of understanding in performance.

The concept of Situation Awareness in **Figure 1** predicates awareness on understanding, consistent with the notion of understanding-based awareness introduced in this section (note that **Figure 1** does not address learning. Accordingly, this article does not expand on relations between learning and understanding, except for the comments in the preceding paragraph). The next section takes a closer look at the process of understanding and provides examples of its successes and failures.

Situational Understanding

Colloquially, "understanding" denotes an ability to figure out what to do when there is no recipe available and no precedent or aid to consult. The dictionary formulation captures the essence of that ability defining understanding

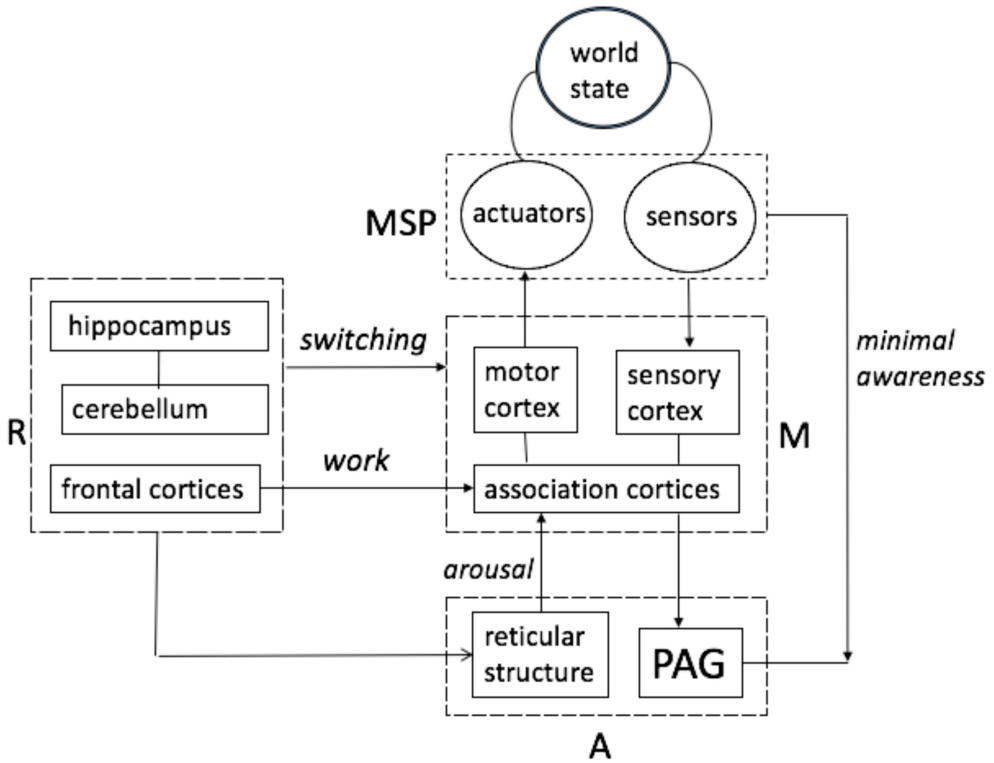


FIGURE 2 | Mental models are structures formed in Memory System (M) and manipulated by Regulatory System (R). Manipulation is enabled by activation (arousal) arriving from the Activation System A (includes Reticular Activating System) and serves to organize activities in Motor-Sensory Periphery (MSP) in such a way that the resulting changes in world states are consistent with the intent originating in R.

(comprehension, grasp) as "apprehending general relations in a multitude of particulars" (Webster's Collegiate Dictionary). In science, relations are expressed by equations. Accordingly, in understanding scientific theory T, apprehending general relations takes the form of "recognizing qualitatively characteristic consequences of T without performing exact calculations" (Criterion for Intelligibility of Theories) (de Regt, 2017, p. 102). The experience of attaining scientific understanding was described by Richard Feynman as having

> "some feel for the character of the solution in different circumstances. . . . if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we "understand" the equation, as applied to the circumstances. A physical understanding is a completely unmathematical, imprecise, and inexact thing, but absolutely necessary for a physicist (Feynman, c/f de Regt, 2017, p. 102).

The Criterion subsumes epistemic and pragmatic aspects of theoretical understanding, i.e., producing explanations of various phenomena and applications in various situations. Figuratively, understanding cuts through the "fog of war" (Clausewitz, 2015/1835) when apprehending battlefield situations and the "fog of mathematics" when apprehending scientific theories.

These notions are consistent with conceptualizations of understanding in psychology [theory of understanding (Piaget, 1975, 1978), theory of fluid and crystallized intelligence (Cattell, 1971, 1978)], emphasizing ability to apprehend relations under novel conditions and in the absence of practice or instruction [fluid intelligence (Cattell, 1971, 1978)]. The term "situational understanding" connotes changing conditions, with situations transforming fluidly into each other (e.g., attack-halt - withdraw - attack. . ., etc.). The remainder of this section presents examples of situational understanding, prefaced by a brief analysis (anatomy of the process) in the next two paragraphs. These examples are followed by preliminary suggestions regarding the underlying mechanisms.

Reduce a multitude of objects to just two, A and B, and consider situation "A moves towards B." In reaching decision that A attacks B, three stages can be identified, with the first one being readily apparent, while the significance of the second is easily overlooked. First, one must perceive A and B, which involves distinct activities (alternating attention between A and B) producing two distinct memory elements (percepts). Second, percept A and percept B must be juxtaposed (grouped), i.e., brought together and held together in memory (call it "working memory"). The task appears to be easy when the activities follow in tight succession (e.g., both A and B are within the field of view) but not so easy when they are separated by large time intervals. The third stage involves establishing a relation, which is predicated on the success of the preceding stages. The second stage is crucial: arguably, the development of understanding was launched by the emergence of mechanisms in the brain

allowing juxtaposition of percepts separated by large stretches of time. At this point, it is informative to note that a recent theory concerning the origins of language capacity in the humans associated this capacity with the emergent availability of mental operation (called Merge) where disjoint units A and B are brought together to produce a new unit (A B) → C amenable to subsequent Merge, (C Q) → Z, and so on (Chomsky, 2007; Berwick and Chomsky, 2016).

Identifying stages in the understanding process helps to appreciate the staggering challenges it faces. When experiencing A, how does the idea of relating A to B come to mind? Figuring out this relation takes effort but the very expression "coming to mind" connotes spontaneity. Accordingly, understanding can break down if the effort is insufficient and/or spontaneous mechanisms fail to deliver. The point is that understanding involves dynamic interplay of deliberate operations and automatic memory processes triggered by the operations that might or might not converge in a grasp. To exemplify the point, consider a syllogism (say, "all humans are mortal, Socrates is a human, therefore Socrates is mortal"). It might appear that the conclusion inescapably follows from the premises but that's an illusion: one might be aware of each of the premises individually but fail to bring them together, and/or the conclusion might either not come to mind or get suppressed upon arrival. Some extreme examples of failed and successful situational understanding are listed next.

On May 17, 1987, the USS Stark on patrol in the Persian Gulf was struck by two Exocet AM-39 cruise missiles fired from an Iraqi F-1 Mirage fighter. An investigation revealed that the aircraft was detected by AWAC (Airborne Warning and Control) patrolling in the area and identified as "friendly." Due to the erroneous initial identification, the captain and crewmembers on the frigate ignored subsequent aircraft maneuvers that were unambiguously hostile (turning, descending and accelerating in the direction of the ship) which resulted in a loss of 37 lives and severe damages to the ship (Miller and Shattuck, 2004).

Between May 9th and June 14th in 1940, France was invaded by the German army. France was one of the major military powers in Europe that maintained adequately equipped forces and, besides, invested tremendous resources in erecting state-of-the-art fortifications on its northern border (the Maginot Line). Despite these preparations, France suffered a historic defeat. Massive literature has been produced over many decades, analyzing the course of events and suggesting various reasons for this colossal and catastrophic failure. A book published in 1941 by a competent French author (served as a liaison officer in the British army during WWI) summarized discussions with French and British officers and political figures before and after the events in question, His analysis offers what appears to be a plausible account and explanations (Maurois, 1941). In particular, the book pointed out that French military and political authorities overestimated the efficiency of the Maginot defenses which stemmed, interestingly, from French technical advances and a sense of engineering superiority. French generals determined Maginot fortifications to be impenetrable on the grounds that they "can be built so rapidly that, in the time necessary for an enemy to take a first line, the defending army can construct a second . . ." (Maurois, 1941, p. 42). A full range of state-of-the-art technologies (reconnaissance photography, advanced communications, etc.) was employed, the terrain was meticulously examined and mapped out and artillery ranges were calculated in advance. "These painstaking labors assured absolute precision of fire. The spotters in front of the forts had before them photographs of the country divided into numbered squares. Perceiving the enemy in square 248-B, all they would have to do was murmur "248-B" into the telephone, and 10 s later the occupied zone would have been deluged with shells and bullets" (Maurois, 1941, p. 48). In short, a confident consensus was predicting that the Maginot fortifications will never be broken through. These predictions turned out to be correct: Germans went around and bypassed the Maginot Line entirely, invading Paris on June 14, 1940.

On January 15th, 2009, the Airbus A320-214 flying from LaGuardia Airport in New York struck a flock of geese during its initial climb out. The plane lost engine power, and ditched in the Hudson River off midtown Manhattan just 6 min after the take off. The pilot in command was Captain Sullenberger (CS), the first officer was Skiles. The bird strike occurred 3 min into the flight and resulted in an immediate and complete loss of thrust from both engines. At that instant, Skiles began going through the three-page emergency procedures checklist and CS took over the controls. In about 30 s, he requested permission for an emergency landing in a nearby airport in New Jersey (NJ) but decided on a different course of action after the permission was granted. Having informed controllers on the ground about his reasons ("We can't do it") and intents ("We're gonna be in the Hudson"), CS proceeded to glide along and then ditch the aircraft in the river. All the 155 people on board survived against a staggeringly bad odds (Suhir, 2019).

The underlying mental process in all three incidents involves item grouping, success or failure in the overall task performance depended on how that step was accomplished. One more example will help to illustrate this contention. Analysis of eye tracking records of ATC controllers revealed latent grouping of aircraft signatures on ATC displays which appeared to be motivated by gestalt criteria (e.g., relative proximity). The probability of detecting possible collision was higher for the aircraft residing in the same group (A B) than for those residing in different groups, (A B) (C D). It was hypothesized that novice controllers could not disable gestalt grouping but the more skilled ones developed a capacity for overcoming its impact on performance (Landry et al., 2001; Yufik and Sheridan, 2002). We now turn to analyzing these examples.

In the USS Stark incident, three items had to be accounted for in the Captain's decision process: *A* = own ship, *B* = AWAC, and *C* = F1 Mirage. In the Captain's mental model, grouping (A B) was the dominant one (i.e., attributing significance to any item C respective A relied entirely on B). The "friendly" determination rendered C irrelevant to A and removed it from consideration. Hence, the "blind spot" on the Iraqi F-1 Mirage fighter whose behavior was displaying signs of attack that could not be any clearer: the aircraft was ascending away from the ship but then sharply changed its course and started descending and accelerating toward the ship.

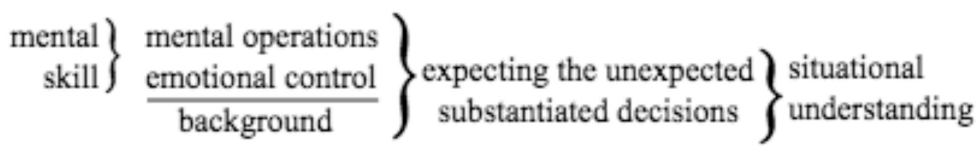


FIGURE 3 | Situational Understanding is a product of background (knowledge, training) and mental skills. A solid horizontal line underscores that skills operate on top of the background.

French military planners recognized the possibilities of German bypassing maneuvers (e.g., attacking through Belgium) but "rationalized them away," i.e., worked out lines of reasoning that rendered them highly unlikely and, ultimately, have forced them out of consideration. French strategic thinking was structured by the experience of trench warfare in WWI when opponents were facing each other from fortified positions and conducted frontal assaults to break through each other's defenses. As a result, the mental models of the leading strategists were focused on the fortifications and defended areas in front of them (A B) while turning a "blind eye " to the adjacent areas (C). Because of the influence earned by the generals in their past victories, these models became the dominant view across the French military, intelligence and political communities. Common sense would suggest that the Maginot Line needed to be "prolonged along the Belgian frontier by fortifications that were perhaps less strong but nevertheless formidable. I received one of the greatest shocks of my life when I saw the pathetic line. . .which was all that separated us from invasion and defeat" (Maurois, 1941, p. 19). The point is that experience-sculpted models can produce pathological tunnel vision which cannot be remedied by reasoning – to the contrary, reasoning confined to the same tunnels can only make them more rigid. Practical validation, an otherwise uncompromisingly reliable criteria, could also do a disservice (one can imagine placing targets in front of the fortifications and, after some extra practice, having them destroyed, not in 10 but in 8 s).

The Airbus A320-214 incident prompted a thorough investigation and analysis that engaged the most advanced investigative and analytic tools available (Suhir, 2012, 2018, 2019; Suhir et al., 2021). Unlike in the previously cited scenarios, this analysis had unlimited access to complete records and could use computer modeling and testing in flight simulators to validate the conclusion. The analysis was centered on probabilistic risk estimates accounting for the human error stemming from imbalances between human capacities (Human Capacity Factor, or HCF) and mental workload (MWL). Ten major contributors into HCF were identified: (1) psychological suitability for the given task, (2) professional qualifications and experience, (3) level, quality, and timeliness of past and recent training, (4) mature (realistic) and independent thinking, (5) performance sustainability (predictability, consistency), (6) ability to concentrate and act in cold blood ("cool demeanor") in hazardous and even in life threatening situations, (7) ability to anticipate ("expecting the unexpected"), (8) ability to operate effectively under pressure, (9) self-control in hazardous situations, and (10) ability to make a substantiated decision in a short period of time. Captain Sullenberger was expected to score high on the majority of these factors. In simulator tests, four pilots were briefed in advance about the entire scenario in full detail and then exposed to simulated conditions immediately after the bird strike. Knowing in advance what to expect, all four were able to land the aircraft. However, when a 30 sec delay was imposed (the time it took Sullenberger to assess the situation and decide on the course of action), all four pilots crashed (Suhir, 2013).

Applying the HCF metric to other examples, it can be suggested that HCF scores reflect capacity for situational understanding, ranging from the bottom low to exceptionally high. For the purposes of this paper, the ten factors can be divided into four groups three of which can be roughly mapped onto components in the architecture in **Figure 2** (roughly, factors 2 and 3 relate to Memory, factors 1, 5, 6 relate to Activation and factor 4 and 9 relate to Regulation) while the forth group is made up of 7, the ability to anticipate ("expecting the unexpected"), and 10, the ability to make a substantiated decision in a short period of time) relate to Situational Understanding, conceptualized here as a product of interplay between the other three groups. **Figure 3** re-phrases this suggestion.

Mental skills operate on top of background, including knowledge and skills acquired in training, but are qualitatively different from those. The distinction extends from responding to unexpected eventualities to constructing scientific proofs or theories where the process of selecting and applying the rules of the theory at each stage cannot be itself governed by another set of rules (de Regt, 2017). In the Airbus incident, emergency rules and training dictated either consulting the emergency checklists or seeking possibilities for heading to the nearest airport. Following either of these courses of action would be both rational (not random or unreasonable) and in line with the cumulative experience in the aviation community, but would have surely killed all on board.

There are five points to be made here. First, an NJ landing was initially considered by CS and implicitly supported by controllers on the ground, as evidenced by the granting landing permission. Second, CS could not even start analyzing the NJ option (i.e., considering the current altitude, airspeed, distance, wind, aircraft characteristics, etc.) but could only develop a "feel" that it would not work out. Third, despite the absence of analysis, the "feel" allowed a substantive decision ("We can't do it"). Forth, having developed the "feel," CS acted on it resolutely, entailing another substantive decision ("We're gonna be in the Hudson"). Fifth,

CS did not know the future but performed comparably to or better than pilots who knew the scenario in advance. In short, CS understood the situation in a process involving three distinct mental operations, as follows.

1. Forceful re-grouping, not derived from any rule or precedent (*jump,* $\curvearrowright$).
   (*A B*) $\curvearrowright$ (*A C*), here A is the aircraft, B is the New Jersey airport and C is the Hudson River.
   The expression reads as "group (A B) *is jumped* to group (A C)."
2. Alternating attention between members inside a group while envisioning variations in their characteristics (*coordination,* $\rightleftharpoons$).
   [*var* (A) $\rightleftharpoons$ *var* (C)], attention alternates between envisioning variations in the aircraft behavior [*var* (A) (e.g., changes in attack angle) and changes along the riverbed (*var* (C)] (e.g., changes in width, curvature, etc.). Reads as "A *is coordinated* with C."
3. Forcefully iterating *coordination* until a particular coordination pattern (*relation)* is apprehended (*blending*, $\leftrightsquigarrow$).
   [(*var* (A) $\leftrightsquigarrow$ *var* (C)], reads as "A *is blended* with C." *Blending* transforms a coordinated group into a cohesive and coherent functional whole so that, e.g., envisioning variations in one member *brings to mind* the corresponding variations in the other one (thinking of ditching near a particular spot brings to mind the required changes in aircraft behavior and, vice versa, envisioning changes in the behavior brings to mind the corresponding changes in the location of the spot). *Blending* establishes *relation R* on the group [*var* (A) $\rightleftharpoons$ *var* (C)] $\rightarrow$ (A *R* C) which gets expressed in substantive decisions ("We're gonna be in the Hudson") and gives rise to probability estimates for coordinated activities ("chances of a successful ditching are not too bad") and their outcomes (more on that in the next section).

Operations *jump*, *coordination,* and *blending* participate in the construction of mental models, culminating in *blending* which makes one *aware of*, i.e., anticipate direct and indirect results of one's actions without considering situational details. Intuitive appreciation of this dualistic relationship between awareness and understanding seems to be the motivation in the Situational Awareness concept and the SA schema in **Figure 1**.

To summarize, understanding involves the construction of mental models that make an adequate performance possible when exploring unknown phenomena and/or dealing with unforeseeable eventualities in the otherwise familiar tasks. In the latter case, understanding enables decision processes that are substantive, short (as compared to the duration of the task), rely on minimal information intake, and achieve results approximating those one would achieve had all the eventualities been known in advance. Importantly, mental models not only generate likelihood estimates for future conditions but make one envision them and then actively regulate motor-sensory activities consistent with the anticipated conditions and in coordination with motor-sensory feedback (hence, the "expecting of the unexpected"). Note that simply to decide or choose immediately requires there to be a space of policies or narratives to select from. The position offered in this paper is that this necessarily entails the ability to represent the (counterfactual) consequences of two or more courses of action—and to select optimally among these representations. What brain mechanisms could underlie this capability?

THE VIRTUAL ASSOCIATIVE NETWORK THEORY OF MENTAL MODELING

The VAN model was motivated by one paramount question ("How does understanding work?") and stems from the three already familiar ideas that can be re-stated as follows:

(1) The world is a stream, and brain processes are dynamically orchestrated to adapt organism's behavior to variations in the stream.
(2) The brain is a physical system, wherein all processes need to be powered by energy extracted from the stream.
(3) Physical systems are dissipative, so any re-organization takes time (instantaneous reorganization would require infinite energy). As a result, adaptive re-organizations are necessarily anticipatory.

Taken together, these ideas entailed the following two hypotheses:

(a) the evolution of biological intelligence has been (selectively) pressured to stabilize energy supplies above some life-sustaining thresholds and
(b) human intelligence was brought about by biophysical processes—discovered by evolution—that allowed for two fundamental mechanisms to emerge: mechanisms that stabilized energy supplies from the outside and those minimized dissipative losses and energy consumption inside the brain. These mechanisms culminate in the uniquely human capacity for understanding, as outlined in the remainder of this section. The next section will suggest a hand-in-glove relationship between thermodynamic efficiency and variational free energy minimization (VFEM) (Friston, 2009, 2010).

Note that the VAN approach is orthogonal to that expressed in the perceptron (neural nets) idea: dynamically orchestrated neuronal structures vs. fixed structures (after the weights are settled), input streams where stimuli combinations are never twice the same vs. recurring inputs. Crucially, accounting for energy and time is integral to the VAN model and alien to the perceptron framework. In short, the VAN and perceptron models reside in different conceptual terrains. The appeal of the former is the possibility of quickly reaching a point where a theory of understanding can be articulated. Technically, the distinction between perceptron and related reinforcement learning and VAN is the distinction between an appeal to the Bellman optimality principle (any part of an optimal path between two configurations

of a dynamical system is itself optimal) and a more generic principle of least action where action corresponds to energy times time. VAN and the free energy principle (a.k.a. active inference) share exactly the same commitments. Note that formulating optimal behavior in terms of a principle of least action necessarily involves time—and the consequences of behavior.

To set the stage, return to **Figure 2**, and think of the world as a succession or stream of states S_i, S_j ... arriving with time interval τ_1, and think of the brain as a pool of N binary neurons. Interaction is driven by the need to extract energy from the world in the amounts sufficient for the pool's survival. Anthropomorphically, this entails recurring cycles of inquiring (What is the current state of affairs in the world?) and forming responses (What shall I do about it?). The sequence of "inquiries" at each cycle can be expanded: What is the state? What can I do about it? What shall I do about? How shall I do it? and so on. Also, different types of neurons can be envisioned and mapped onto different components in the architecture in **Figure 2** (sensory neurons, motor neurons, etc.).

Whatever the composition of the pool and the content and order of the inquiries, activities in the pool boil down to selectively flipping (exciting or inhibiting) neurons in a particular order. Make two assumptions: (a) each state S_i can emit energy reward $_i$ ranging from 0 to some maximum $\triangle_i^{max}$, depending on the order and composition of "flippings" in the pool and (b) each "flip" consumes energy d (at the first approximation, let all flips be powered by the same energy amount). The problem facing the pool can be defined now as maximizing energy inflows while minimizing the number of flips. It will be argued, in four steps, that understanding involves a particular strategy for satisfying this dual objective (step 4 defines architecture for understanding).

Step 1. Neuronal Groupings

A pool of N binary neurons admits 2^N configurations so that, in principle, selecting a rewarding configuration for a particular world state can pose a problem that grows in complexity with the size of the pool (associating complexity measure with the number of options). The problem is alleviated when choices are dictated by the world state itself (i.e., each stimulus in the composition of S_i excites particular neurons) but, otherwise, the pool needs to choose between 2^N options.

Assume that a mechanism exists to partition the pool into m groupings [call them neuronal assemblies (Hebb, 1949, 1980)] such that all the neurons in every group behave in unison. Such partitioning would offer more efficiency, reducing the number of choices from 2^N to 2^m. The remedy is radical because it not only puts a lid on complexity growth but causes complexity to decrease steeply with the size of the pool (e.g., partitioning pool of 10 neurons into 5 groups yields 2^5: 1 reduction in the number of options while having 5 groups in a pool of 100 neurons obtains 2^{95}: 1 reduction). Complexity reduction translates into an increase decision speedup (e.g., equating complexity to time-complexity, by assuming one choice per unit time) and internal energy savings. Indeed, complexity reduction can be regarded as underlying all (i.e., universal) computation; in the sense of algorithmic complexity and Solomonov induction. The benefits of compression and complexity minimization come at a price:

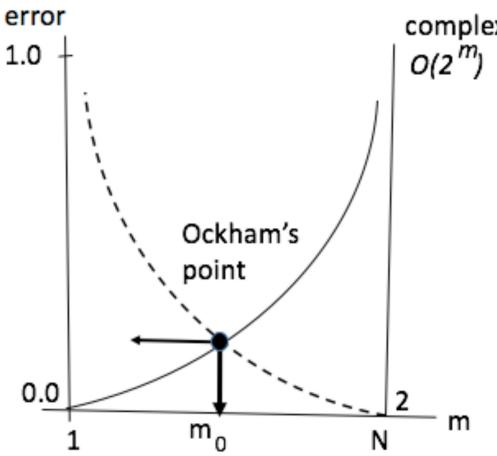


FIGURE 4 | Self-partitioning in the neuronal pool radically impacts pool's capacities in responding to world streams and involves trade-offs between time and accuracy, as a function of group size. The relationship is non-linear, creating long tail areas where, on the one side, sacrificing speed (increasing the number of groups) produces no appreciable improvements in accuracy ("useless details") and, on the other side, small speed gains produce quickly increasing errors ("useless generalities"). A narrow inflection zone (Ockham's point, or *O*-point) lies between the tail areas.

imploding complexity is accompanied by exploding error — as the loss of degrees of freedom precludes an accurate prediction. This trade-off between accuracy and complexity is illustrated in the notional diagram in **Figure 4** (error η_i is measured by the difference between energy gain $\triangle_i^N$ obtainable in the pool without partitioning and gain $\triangle_i^m$ yielded by m- partitioning).

To illustrate, veering to the left of the *O*-point in the Airbus accident would be akin to CS receiving advice "aviate, navigate, communicate" from the ground controllers, which is a paramount principle in aviation human factors (Wiener and Nagler, 1988) but hardly a useful guidance under the circumstances, while veering to the right would be like offering a refreshment course in plane aerodynamics. Depending on the task, the relative width of Ockham's zone on the group size axis can be very small so the ability to stay within it (e.g., not going through emergency checklists, discontinuing communications, etc.) can make vital differences in the performance outcomes. Put simply, there is a right level of "grouping" or "course graining" that provides the right balance between accuracy and complexity. Statistically speaking, this corresponds to maximizing marginal likelihood or model evidence.

Arguably, the emergence of grouping mechanisms in the neuronal substrate was a major discovery in the evolution of biological intelligence (from sensing to understanding). Accordingly, the concept of neuronal assembly remains the single, most revealing idea at the foundation of neuroscience (Hebb, 1949, 1980). Neuronal grouping opened new avenues for development, *via* fine-tuning and manipulation of the groups. Pursuing such adaptive improvements equates to bending curves and "pushing" the Ockham's point toward obtaining minimal error in the smallest number of groups (see **Figure 3**). It was

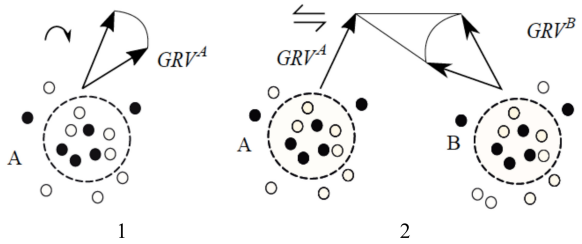


FIGURE 5 | (1) Patterns of excitation-inhibition within groups can be varied, which can be expressed as rotation of group response vectors (symbol ↷ denotes operation "rotation of group response vector"). **(2)** Movements of group response vectors can be coordinated: every position of GRV^A determines a range of admissible positions for GRV^B, and vice versa, movement of one vector causes re-positioning of the other one (i.e., thinking of changes in A *brings to mind* the corresponding changes in B).

subsequently argued that thermodynamics has been doing the "pushing" (Yufik, 1998, 2013), we will touch on that later.

Step 2. Varying and Coordinating Group Activities

On-off decisions on neuronal groups can be dynamically nuanced to allow more close tracking of the world stream, by, first, tuning receptive fields in individual neurons and, second, by varying excitation–inhibition balance within each. A convenient expression of that strategy can be obtained by summing up response vectors of all the participating neurons in a group to obtain "group response vector" (GRV) and then characterizing activity variations inside a group as patterns in the movement of GRV. Finally, mechanisms for inter-group coordination would develop on top of the mechanisms for controlling intra-group variations. Coordination involves mutual constraints, i.e., variations in one group can both trigger and limit the range of variations in another one. Mutual constraints reduce the number of options, thus shifting the *O*-point down and to the left. **Figure 5** depicts progression from intra-group variation to inter-group coordination.

One of the cornerstone findings in neuroscience revealed that movement control (e.g., extending hand toward a target) involves a rotation of response vectors in groups of motor neurons, as in **Figure 5.1** (Georgopoulos and Massey, 1987; Georgopoulos et al., 1988, 1989, 1993). Accordingly, complex coordinated movements can involve coordinated rotation of group response vectors in synergistic structures in the motor cortex comprising multiple neuronal groups (Latash, 2008).

Step 3. Neuronal Packets and Brain Energy Landscapes

The following hypotheses is central in the VAN model: neuronal assemblies are formed as a result of phase transitions (Kozma et al., 2005; Berry et al., 2018) in associative networks, when tightly associated subnets become separated by energy barriers from their surrounds (c.f., the formation of droplets in oversaturated vapors). The term "neuronal packet" was coined in Yufik (1998) to denote neuronal assemblies bounded by energy barriers. It can be argued that Hebb's insight recognizing assemblies as functional units in the nervous system (as opposed to attributing this role to individual neurons) necessarily implied the existence of biophysical mechanisms that keep such assemblies together, separate them from the surrounding network and make it possible to manipulate them without violating their integrity and separation. On that argument, the VAN theory only makes explicit what was already implied in the idea of neuronal assembly. **Figure 6** elaborates this contention.

Associating boundary energy barriers with biological neuronal groups expresses a non-negotiable mandate that operations on such groups, including accessing the neurons inside, varying excitation-inhibition patterns in the groups, removing neurons from a group, etc. all involve work and thus require a focused energy supply to the group's vicinity sufficient for performing that work. Multiple packets establish an energy landscape in the associative network, as shown in **Figure 7**.

Packets are internally cohesive and externally weakly coupled (i.e., neurons in a packet are strongly connected with each other and weakly connected with the neurons in other packets), the cohesion/coupling ratio in a packet determines the depth of energy "well" in which it resides: the deeper the well, the more stable the packet, which translates into reduced amounts of processing and higher degree of subjective confidence when packet contents are matched against the stream [packets respond to correlated stimuli groupings, the number of matches sufficient for confidently identifying the current input decreases as the cohesion/coupling ratio increase (Malhotra and Yufik, 1999)]. Changes in the landscape, as in **Figure 6**, result from changes in arousal accompanying changes in subjective values (importance) attributed to the input (objects, situation): the higher the value attribute to an object, the deeper the corresponding well becomes (more on that shortly).

The notions of neuronal packets and energy landscape in Yufik (1996, 1998); Yufik and Sheridan (2002) anticipated experimental and theoretical investigations of cortical energy landscapes (Watanabe et al., 2014; Gu et al., 2017, 2018; Kang et al., 2019). However, packet energy barriers are amenable to direct experience, as was first intimated by William James in his classic "The Principles of Psychology" back in 1890, as follows. To access an item in memory, one must make attention

> "linger over those which seem pertinent, and ignore the rest. Through this hovering of the attention in the neighborhood of the desired object, the accumulation of associates become so great that the combined tensions of their neural processes break through the bar, and the nervous wave pours into the track which has so long been awaiting its advent" (James, 1950/1890, v. 1, p. 586).

To appreciate the insightful metaphor "breaking through the bar," think of desperately trying to recollect the name of an acquaintance that escaped you just at the moment you were making an introduction. With a stunning insight and vividness (**Figure 8**), James describes the experience of mounting mental effort to access packet's internals from the surrounding associative structure:

> "Call the forgotten thing Z, the first facts with which we felt it was related, a, b, c and the details finally operative in calling it up, l,

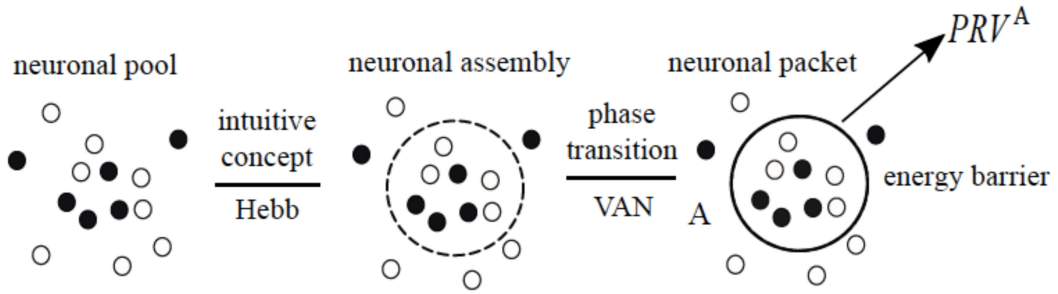


FIGURE 6 | The idea of assembly expresses the notion that groups of tightly associated neurons form cohesive units distinct from their surrounds in the network (associative links are not shown). The notion of a neuronal packet expresses, in the most general terms, a mechanism for forming and stabilizing such units in a material substrate (i.e., phase transition and emergence of an energy barrier in the interface between the phases). PRV^A denotes "packet response vector."

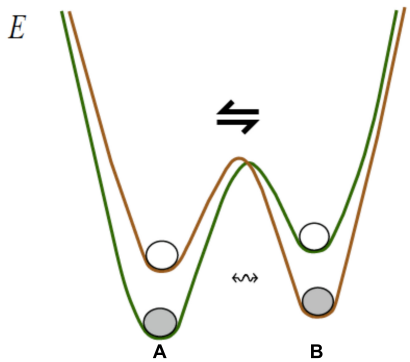


FIGURE 7 | Associative structures reside in continuous energy landscape. Coordinating objects A and B occupying different minima (A B) → (A ⇋ B) requires repetitive climbing over the energy "hill" between the minima. Deformation in the landscape (lowering the "hill") enables *blending* (A ⇋ B) → (A ⇜⇝ B), producing a structure where A and B remain distinct and, at the same time, capable of constraining each other's behavior.

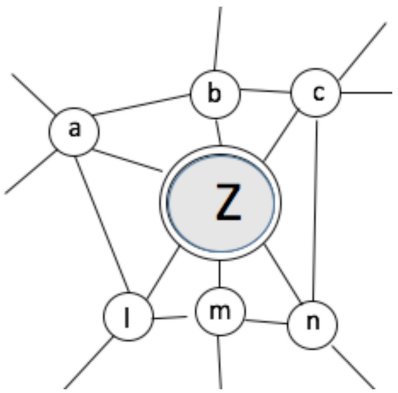


FIGURE 8 | Accessing contents of packet Z requires sustained attention in the associative neighborhood until effort is mounted sufficient for overcoming "resistance"(i.e., boundary energy barrier) (adopted from James, 1950/1890, v. 1, p. 586).

> m and l. . . .The activity in Z will at first be a mere tension, but as the activities in a, b and c little by little irradiate into l, m, n, and as all these processes are somehow connected with Z, their combined irradiation upon Z . . .succeed in helping the tension there to overcome the resistance, and in rousing Z to full activity" (James, 1950/1890, v. 1, p. 586).

Building on the notions in **Figure 7**, assume, first, that Z admits a number of distinguishable states $Z = Z_1, Z_2, \ldots, Z_k$, second, another packet $Q = Q_1, Q_2, \ldots, Q_m$ exists somewhere in the associative network, third, attention alternates between varying states in Z $Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Z_k$ and Q $Q_1 \rightarrow Q_2 \rightarrow \ldots \rightarrow Q_m$ (i.e., rotating packet vectors) until, finally, a particular form of coordination between the variation patterns is established (relation r), producing a coordinated relational structure Z *r* Q. With that, a model is formed expressing variations in the world stream in terms of objects, their behavior and inter-object relations (more on that shortly). Transporting James' vivid account into modern context, "hovering of the attention" can be compared to burning fuel in a helicopter hovering over a particular spot, and inter-packet coordination is like keeping two helicopters airborne and executing different but coordinated flight patterns. Finally, forcing changes in the landscape and establishing coordination, as in **Figure 6**, is analogous to letting the helicopters roll on the ground and having them connected by a rod to coordinate their moves. The following two suggestions reiterate these notions more precisely.

First, alternating between the packets is an effortful process critically dependent on the strength of "resistance" offered by the energy barrier: excessive height will make the packets mutually inaccessible while low barriers will make them less stable and thus disallow sustained and reproducible variations. In short, the process is contingent on maintaining a near-optimal height of energy barriers throughout the landscape, as suggested in **Figure 6**.

Second, establishing relations replaces effortful alternations between packets with effortless (automatic) "facilitation" (the term is due to Hebb, 1949). Stated differently, a rule "varied together, coordinated together" can be suggested as a complement to Hebb's "fire together, wire together" rule, extending its application from neurons to packets. Facilitation underlies the experience of *coming to mind* when thinking of changes in Z *brings to mind* the corresponding changes in Q,

as in **Figure 5.2**. More generally, packets become organized (blended) into a model yielding the capacity to "have some feel for the character of the solution ….without actually solving the equations" (Feynman, see section "The Virtual Associative Network Theory of Mental Modeling"). Stated differently, one becomes aware of the direction in which changes in one model component impact behavior of the other ones and of the entire composition, consistent with the insight expressed in **Figure 1**. Situational "feeling" is coextensive with reaching understanding and obtaining complexity reduction in the modeling process on a scale ranging from small in simple situations to astronomical in complex ones.

To appreciate the significance of the benefit, think of a most rudimentary task, e.g., a chimpanzee connecting sticks and climbing on top of piled boxes to reach some fruit. Connecting sticks involves trying out different random variations until the proper coordination is encountered (Koehler, 1999). Connected sticks become a physical unit that can be physically coordinated with other units (i.e., carried on top the boxes) which is contingent on forming and coordinating the corresponding memory units (pairwise coordinations, i. e, stick1- fruit, stick2-fruit, box1- fruit, etc. might never amount to a solution). Ability to temporarily decouple mental operations from their motor-sensory expressions and to combine coordinated packets into stable functional units amenable to further coordination (that is, the ability to think and understand) separates humans from other species. Piaget articulated these notions convincingly, by pointing at the "contrast between step-by-step material coordinations and co-instantaneous mental coordinations" and demonstrating in multiple experiments that "mental co-ordinations succeed in combining all the multifarious data and successive data into an overall, simultaneous picture which vastly multiplies their power of spatio-temporal extensions. . ." (Piaget, 1978, p. 218).

Step 4. Architecture for Understanding

Figure 9 positions mechanisms of packet manipulation in the three-partite brain architecture in **Figure 2** superposed on the schema of Situational Understanding in **Figure 3**.

Two blocks are identified, denoting two classes of memory processes and operations: block A is shared across many species while block B is exclusively human, as follows. Block A limits memory processes to the formation of associative networks and packets, and allows for the rotation of packet vectors. Block B allows other operations leading to construction of mental models and operations on them.

Block A (block B is absent or underdeveloped) reflects cognitive capacities in non-humans, from simple organisms to advanced animals. Rudimentary forms of learning reduced to selective formation and strengthening of associative links are available in simple organisms (e.g., worms, frogs) and decorticated animals [e.g., rats having 99.8% of the cortex surgically removed (Oakley, 1981)]. Intact rodents occupy an intermediate position in the capacity ladder [learning involves formation of a few neuronal groups that get selectively re-combined depending on changes in the situation (Lin et al., 2006)]. Apes and some avians can learn to coordinate a few objects (link C).

Block A operates on the associative and packet network in block B while leaving the mosaic of associative links intact. Flexible neuronal "maneuvers" [fluid intelligence (Cattell, 1971, 1978)] underlie management of competing goals and other executive functions (Mansouri et al., 2009, 2017) and involve selective re-combination of packets, producing a hierarchy of relational models (hierarchy of flexible relational structures developing on top of an associative network is called *virtual associative network)*. Interactions between levels are two-directional, with the top-down processes selectively engaging lower levels, down to deployment of sensorimotor resources which can entail changes in the bottom associative network due to sensorimotor feedback (please see below).

Link D places energy distribution across the packet network under regulatory influence (volitional control), thus making it an integral part of a human cognitive system, as suggested in **Figure 10**.

In the extreme, low barriers allow floods of irrelevant associations while high barriers confine attention to a few familiar associations. Accordingly, optimal arousal obtains optimal task space partitioning (m_0) yielding optimal performance.

Arousal-induced changes in the landscape account for the levels of awareness, from vegetative wakefulness (flat landscape) to understanding-based awareness (optimal landscape, see **Figure 1**). Subjective experience of arousal varies from fear, stress, anxiety on the one and of the spectrum to excitement and exhilaration on the other end. Accordingly, moving along the spectrum changes the topological characteristics of the energy landscape: from fragmented access (i.e., some areas are inaccessible) to the unrestricted accessibility of a flat surface. Stress-induced changes in landscape topology are likely to underlie the idea of "suppressed memories" treated in psychoanalysis (disturbing memories are not erased or degraded but become "walled off " behind high barriers, so access to them can be restored if the barriers are lowered. Treatment that concentrates on the associative neighborhood (see **Figure 6**), as in dream analysis, seems to be appropriate for that purpose). Methods of memory recovery were disputed, on the grounds that it might be as likely to conjure false memories as to recover access to the lost ones (Loftus and Ketcham, 1996). However, creation and suppression are two sides of the same coin, i.e., the same mechanism that facilitates creative re-combination of memory structures can block access to some of them. Stress-induced landscape distortions can be responsible for other psychological symptoms, such as obsessive thoughts.

We will now return to the examples above, this time applying the notions of the VAN framework. The USS Stark incident and the Maginot catastrophe were not a product of insufficient training or illogical reasoning but resulted from understanding failure, that is, the inability to form "mental co-ordinations . . . combining all the multifarious data and successive data into an overall, simultaneous picture" (Piaget, 1978, p. 218). Despite differences in circumstances, the nature of cognitive deficiency was the same in both scenarios: an inability to overcome the resistance of elevated energy barriers, which resulted in

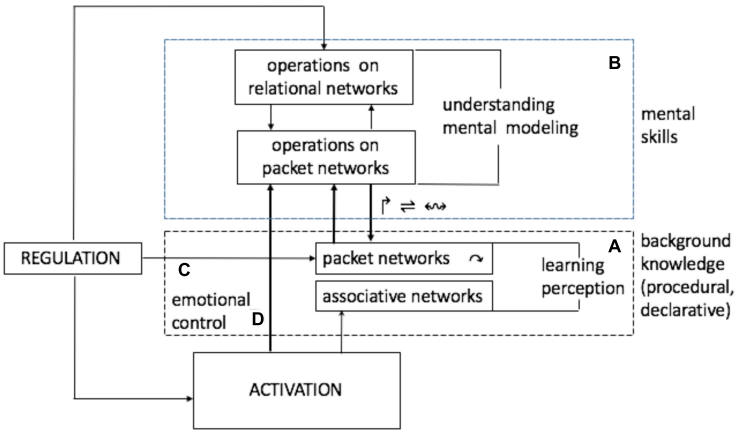


FIGURE 9 | Architectures for understanding. This diagram represents cognition as a regulatory process that is directed at adapting (matching) behavior variations in the organism to condition variations in the world stream and is powered by energy inflows extracted from the stream. Organization in the system comprises different structures submitted to regulation [from tuning receptive fields in individual neurons (Fritz et al., 2003, 2005), to rotating packet vectors, to constructing and manipulating mental models], seeking to stabilize energy inflows while minimizing metabolic costs.

fragmented (as opposed to simultaneous) "pictures." On the VAN theory account, the captain's mental model in the first scenario comprised two uncoordinated packets: *A* = ship, objects relevant to the ship, and *B* = all other objects. A highly valued but erroneous AWAC classification placed the Iraq jet in the second group, and the captain's mental skills did not allow crossing the A | B barrier and coordinating members of B with members of A. In the second scenario, mental model of the high command comprised *A* = fortifications, defended area in front of fortifications and *B* = adjacent areas, objects in the adjacent areas separated by an energy barrier that turned out to be insurmountable due to overvalued significance of past experiences. French high command, as a collective entity, demonstrated low level of self-control under fear and anxiety brought about by the anticipated German attack, which caused them to fall back on the past tactics and made them "fanatically uninterested" in deviating from them.

By contrast, a high degree of self control ("ability to operate effectively under pressure, self-control in hazardous situations" Suhir, 2013) demonstrated in the Airbus incident made possible suppressing fear and bringing arousal to a level enabling situational understanding manifested in overcoming the inertia of training and customary practices (regulations, authority of the ground control, etc.), "feeling" the appropriate course of action, and making decisions at a substantive level ("we can't do it," "we're gonna be in the Hudson"). The well coordinated mental model regulated subsequent activities in a top-down fashion, by selectively engaging skills and knowledge in the pilot's background repertoire as necessary for coordinating flight pattern with river characteristics to enable a safe ditching. **Figure 11** depicts a succession of mental operations.

Figure 10 underscores that mental models are regulatory structures that, beside supplying "pictures," control their own execution *via* dynamic coordination of various data streams in the motor-sensory loop completed *via* environmental feedback [sensory streams include visual input (e.g., river shape), motor-kinesthetic input, etc.]. Execution is accompanied by a feeling of confidence in reaching the objective (e.g., safe ditching) that varies depending on the varying degree of correspondence between the envisioned outcomes of control actions and the actually observed ones. Technically, grasp can be said to establish a functional on the space of packet vectors that returns confidence values for different patterns of inter-packet

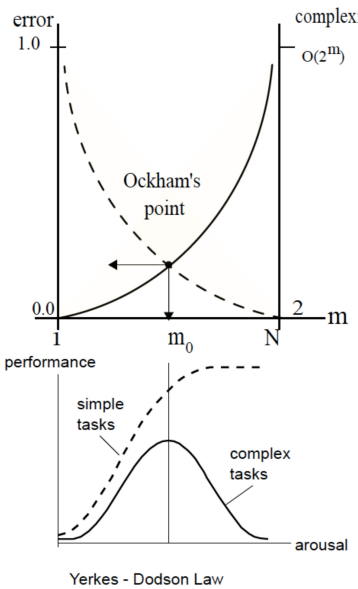


FIGURE 10 | The shape of the energy landscape is a function of interplay between arousal and value distribution across the packets (reflecting value distribution in the corresponding objects). Heightened arousal lowers energy barriers across the landscape enabling coordination of distant packets, as might be necessary for unfamiliar and complex (creative) tasks, while decreasing arousal elevates the barriers thus restricting coordination to proximal packets (which might suffice for simple and familiar tasks).

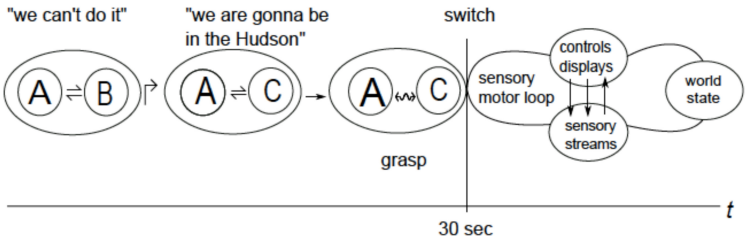


FIGURE 11 | Here A - aircraft, B - New Jersey airport, C - Hudson River. Mental operations are accompanied by imagery and remain decoupled from the motor-sensory feedback until, following grasp, the motor-sensory system gets engaged.

coordination. Behavior of the functional depends on the vector space topology, i.e., accessibility between packets.

Following grasp, the repetitive successful exercise of a newly formed model causes its stabilization, which is captured, to an extent, in the concept of frame (schema, script, etc.) defined as a fixed memory arrangement comprised of components (slots) with variable contents (e.g., script of visiting a restaurant comprises slots "entering," "being seated," "studying menu," etc. (Schank and Abelson, 1977; Norman, 1988). A few comments on the frame idea are offered in the discussion part.

Since the VAN theory pivots on the notion of energy efficiency in the brain, a brief excursion into that subject is in order. The notion that neuronal system optimizes energy processes (Yufik, 1998, 2013; Yufik and Sheridan, 2002) is consistent with later theoretical proposals (e.g., Niven and Laughlin, 2008; Vergara et al., 2019; Pepperell, 2020, 2018) and an increasing number of experimental findings (the discussion section offers a brief review of some data). To appreciate the sources of energy efficiency inherent in the VAN concept, consider the following. In an associative network, excitation in any node or group of nodes can propagate throughout the entire network. By contrast, propagation of excitations induced within a packet is obstructed by boundary energy barriers (i.e., crossing a barrier incurs energy costs). Moreover, seeking further energy savings drives the system toward constraining intra-packet activities to packet subsets and, when crossing the barriers, to engage only packets amenable to mutual coordination. In this way, formation of mental models comprising entities (packets), behavior (transition between intra-packet activity patterns) and relations (inter-packet coordination) expresses the dual tendency to increase the efficacy of action plans (enabled by situation understanding) while decreasing the costs of such planning. A reference to neuronal processes that might be responsible for some of these phenomena will conclude this section.

Interaction between neuronal cells is mediated by several types of substances, including neurotransmitters and neuromodulators. Neurotransmitters act strictly locally, i.e., they are released by a pre-synaptic neuron and facilitate (or inhibit) generation of action potentials in a single post-synaptic target. By contrast, neuromodulators act diffusely, i.e., they are released to a neighborhood as opposed to a specific synapse and affect a population of neurons in that neighborhood possessing a particular receptor type (metabotropic receptors). Neuromodulators control the number of neurotransmitters synthesized and released by the neurons, thus allowing up- or down- regulation of interaction intensity. Neurotransmitters move through fast-acting receptors metabotropic receptors are slow-acting receptors that modulate the functioning of the neuron over longer periods (Avery and Krichmar, 2017; Pedrosa and Clopath, 2017). Neuromodulators were found to provide emotional content to sensory inputs, such as feelings of risk, reward, novelty, effort and, perhaps, other feelings in the arousal spectrum (Nadim and Bucher, 2014). It can be suggested that James' vivid depiction of "hovering of the attention in the neighborhood of the desired object" provides an accurate introspective account of the work invested in regulating neuromodulator concentration and neurotransmitter production at the packet boundary, which amounts to lowering the energy barrier until "the combined tensions of neural processes break through the bar" (James, 1950/1890, v. 1, p. 586). Since neuromodulators are slow acting, the packet remains accessible for a period of time sufficient for the task at hand.

To summarize, psychology usually treats awareness as a necessary but insufficient prerequisite for reaching understanding (e.g., one can be fully aware of all the pieces and their positions on the chessboard but fails to understand the situation). According to the present theory, predicating situation awareness on situation understanding, as in **Figure 1**, refers to *understanding-based awareness* (see section "Levels of Awareness") and expresses a keen insight consistent with one of the key assertions in the VAN theory: the experience of attaining understanding accompanies emergence of a synergistic (coherent and cohesive) mental models, simulating (envisioning) possible actions on particular elements in such models generates awareness of the constraints and likely consequences of those actions in the other elements throughout the model (hence, the *situation awareness*).

INTEGRATING VIRTUAL ASSOCIATIVE NETWORK INTO THE VARIATIONAL FREE ENERGY MINIMIZATION FRAMEWORK

The Free Energy Minimization principle offers a "rough guide to the brain" (Friston, 2009) and extends to any biological system, from single-cell organisms to social networks (Friston, 2010).

The central tenets of the VFEM come from the realization that any living system must resist tendencies to disorder, including those emanating from the environment, while obtaining means for resistance from that same environment:

> "The motivation for the free-energy principle . . . rests upon the fact that self-organizing biological agents resist a tendency to disorder and therefore minimize the entropy of their sensory states " (Friston, 2010, p. 293).

The success or failure of the enterprise depend on the system's ability to adapt, *via* forming models of the world used to predict the forthcoming conditions. The VFEM principle expresses this insight in information-theoretic terms, *via* the notion of variational free energy defined as follows:

> "Free-energy is an information theory quantity that bounds the evidence for a model of data . . . Here, the data are sensory inputs and the model is encoded by the brain. . . .. In fact, under simplifying assumptions. . .it is just the amount of prediction error" (Friston, 2010, p. 293).

Technically, variational free energy is F_v is defined as surprise (or self-information) - ln p (y| m) associated with observation y under model m, plus the difference between the expected and the actual observations (i.e., the prediction error under model *m*), measured as a Kullback-Leibler divergence D_{KL}, or entropy, quantifying distinguishability of two probability distributions.

This section adopts the simplifying assumptions and equates variational free energy to prediction error. The VFEM principle conceptualizes minimization of prediction error as a causal factor guiding interaction with the environment, as follows:

> "We are open systems in exchange with the environment; the environment acts on us to produce sensory impressions and we act on the environment to change its states. This exchange rests upon sensory and effector organs (like photoreceptors and oculomotor muscles). If we change the environment or our relationship to it, sensory input changes. Therefore, action can reduce free-energy (i.e., prediction errors) by changing sensory input, whereas perception reduces free-energy by changing predictions" (Friston, 2010, p. 295).

Adaptive capacities culminate in the ability to adjust accuracy, or precision to optimally match the amplitude of prediction errors, as follows:

> "Conceptually, precision is a key determinant of free energy minimization and the enabling – or activation – of prediction errors. In other words, *precision determines which prediction errors are selected* and, ultimately, how we represent the world and our actions upon it. . . ..it is evident that there are three ways to reduce free energy or prediction error. First, one can act to change sensations, so they match predictions (i.e., action). Second, one can change internal representations to produce a better prediction (i.e., perception). Finally, one can adjust the precision to optimally match the amplitude of prediction errors" (Solms and Friston, 2018).

The VAN theory instantiates the VFEM principle for the human brain, identifying understanding with a particular strategy for predictive error reduction and a particular form of precision adjustment. In this way, the VAN theory proposes some substantive contributions to the VFEM framework, including the following. Firstly, the VFEM principle envisions changing actions to change sensations and changing internal representations in order to change perceptions. The VAN theory envisions, in addition, changes in the internal models to produce and change understanding. Secondly, "the motivation for the free-energy principle . . .. rests upon the fact that self-organizing biological agents resist a tendency to disorder and therefore minimize the entropy of their sensory states" (Friston, 2010, p. 293). VAN postulates that self-directed construction of mental models constitutes a form of self-organization in the brain that reduces the entropy of its internal states (Yufik, 2013, 2019) (more on that important point in the next section). Thirdly, according to the VFEM, error minimization brings about the minimization of energy consumption in the brain. By contrast, VAN attributes ontological primacy to energy processes and derives error reduction from the pressure to reduce energy consumption.

Technically, the VAN and VFT share the same commitment to finding the right balance between accuracy and complexity, i.e., the right kind of grouping or course graining that conforms to Occam's principle. This follows because variational free energy is a bound upon the log of marginal likelihood or model evidence (i.e., negative surprise or self information). As noted above, the marginal likelihood can always be decomposed into accuracy and complexity. This means that the energy landscapes above map gracefully to the variational free energy landscapes that attend the free energy principle. The link between the informational imperatives for minimizing prediction errors and the thermodynamic imperatives for efficient processing rest upon the complexity cost, that can be expressed in terms of a thermodynamic cost (*via* the Jarzynski equality). An example will illustrate the underlying notion of efficiency from both a statistical and thermodynamic perspective:

Consider a frog trying to catch flies and getting disappointed by the results (too many misses). To secure a better energy supply, the frog can start shooting its tongue faster, more often, etc. If the hit/miss ratio does not improve and the frog keeps shooting the tongue in vain, it will soon sense the amplitude of prediction error unambiguously – by dying from exhaustion. Presume that neuronal mechanisms emerge that improve the score by improving sensory- motor coordination. In principle, this line of improvement could continue indefinitely making the frog progressively more sophisticated hunter, except that the mechanisms can require more neurons engaged in more intense activities which will result in increasing energy demands that can outweigh increases in the intake (besides, there are obvious physiological and physical limitations on the brain size, and neither neurons can become smaller, nor the underlying chemical processes can run faster).

Consequently, radical behavior improvements are predicated on discovering mechanisms that deliver them without increases in the size of neuronal pool and/or neuronal activities, that is, without increases in internal energy consumption or, better yet, entailing energy savings. The point is that such mechanisms might or might not emerge, and error reduction is a consequence of their development, as opposed to such mechanisms being a guaranteed accompaniment of error reduction. With these

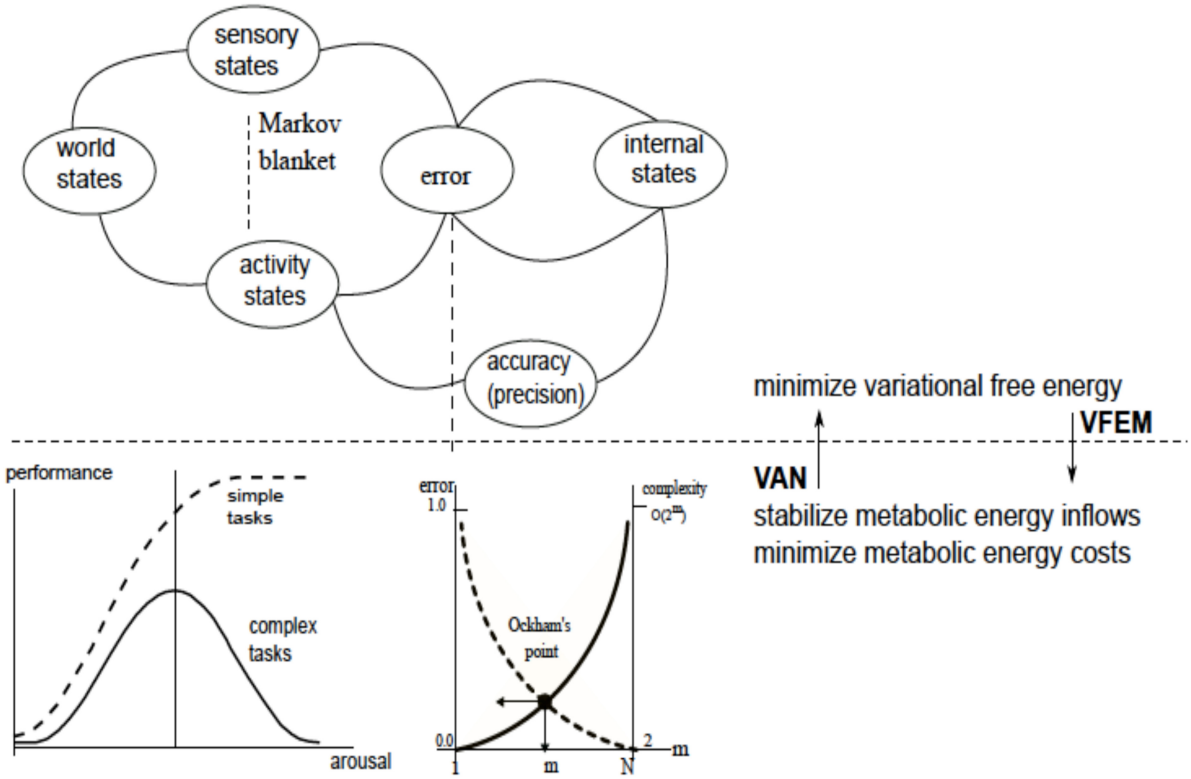


FIGURE 12 | VAN theory instantiates VFEM framework, by accounting for error optimization mechanisms underlying human understanding. The figure above the horizontal dotted line depicts a VFEM construct where adaptive interaction between world states and the brain (internal states) is conducted *via* motor-sensory loop and pivots on the mechanisms of error minimization and precision adjustment (adopted from Solms and Friston, 2018). Figure below the horizontal dotted line summarizes the VAN approach deriving error optimization from capabilities inherent in self-directed packet manipulation (see **Figure 9**). Vertical dotted lines suggest mapping between the VFEM and VAN constructs. The sensory and activity states constitute Markov Blanket that shields internal states from the world states and, at the same, mediates interaction between them.

caveats, **Figure 12** suggests a straightforward integration of the VAN model into the VFEM framework.

MACHINE SITUATIONAL UNDERSTANDING

This section illustrates the function of machine situational understanding and discusses approaches toward its implementation.

Machine Understanding

A machine can be said to possess situational understanding to the extent it can:

(a) accept task definition from the operator expressed in substantive terms,

(b) evaluate a novel, unfamiliar situation and develop a course of action consistent with the task and situational constraints (the available time, data sources, etc.) and

(c) communicate its decisions and their reasons to the operator in substantive terms.

In other words, decision aid is attributed a degree of situational understanding if the operator feels that the machine input contributes into his/her situational awareness and can be sufficiently trusted to adjust his/her own situational understanding and to act on machine advice. In the VAN framework, substantive expressions address objects (entities), their behavior, and forms of behavior co-ordination (relations). The same three examples will illustrate these suggestions.

In the USS Stark incident, an on-board situation understanding aid (SUA) could overrule AWAC target classification and issue a warning like "Attention: there is 0.92 probability that this is enemy aircraft." The chances that the warning will be trusted and acted upon will improve significantly if, when asked "How do you know?" the system would reply with "The aircraft was ascending but then turned sharply and started descending and accelerating toward you." Assuming that the captain interacts with the ship systems *via* the aid, the SUA would accept the captain's command "Engage the target" and initiate activities by the engagement protocol [note that learning systems (e.g., deep learning) are capable of reliably detecting and identifying objects but are limited in their ability to apprehend relations and explain their decisions to the user].

In the Airbus incident, the SUA could be tasked with interacting with ground control to request permission to land in NJA, and could respond with "We are not going to make

it." Improving situation understanding in the Maginot scenario would require breaking a rigid mental template, some (tentative) suggestions for a possible role of SUA will be made shortly, after introducing VAN computational framework.

Virtual Associative Network Computational Framework and Virtual Associative Network/Variational Free Energy Minimization Integration

The VAN computational framework was dubbed "gnostron" (Yufik, 2018), to underscore distinction from "perceptron": perceptron has a fixed neuronal structure while gnostron is a neuronal pool where structure evolves gradually and remains flexible. Gnostron formalism is a straightforward expression of VAN considerations summarized in section "Integrating Virtual Associative Network into the Variational Free Energy Minimization Framework," as follows.

World is a stream of stimuli $S = s_1, s_2, \ldots, s_M$ arriving in different combinations at a pool comprised of N neurons $X = x_1, x_2, x_N$, with each neuron responding probabilistically to a subset of stimuli. In turn, the stimuli respond probabilistically to the neurons that pool mobilizes and "fires at" them, by releasing energy deposits (neuronx_i has receptive field $\mu_{ij}, \mu_{ih}, \ldots \mu_{ik}$, here μ_{ih} denotes probability that stimulus s_h will release deposit E_h in response to the pool having fired x_i). Mobilization (selecting neurons and preparing them to fire) takes time and neurons, after having fired, need to time to recover, which forces the pool to engage in anticipatory mobilization. Engaging x_i consumes energy δ_i comprising the work of mobilization ρ_i and the work of firing ν_i, $\delta_i = \rho_i + \nu_i$ (note that mental operations are constituents of mobilization).

The pool's survival depends on maintaining net energy inflows (cumulative deposits minus cumulative expenditures) above some minimal threshold, which includes the requirement that the average mobilization period is commensurate with the tempo of stimuli arrival. This formulation translates the problem of survival and adaptive efficiency into that of probabilistic resource optimization: orchestrate the pool's activities (mobilization, firing and inhibition, or demobilization) consistent with variations in the stream so that energy inflows are maximized (or stabilized at some acceptable level) while energy expenditures are minimized. Conceptualizing cognitive processes as dynamic optimization of neuronal resources (Yufik, 1996, 1998) is consistent with the recent views associating advanced cognitive functions with the ability to monitor the significance of multiple goals and flexibly switch between them so that the rewards yielded by the goals are maximized and the associated neuronal costs are minimized (e.g. Mansouri et al., 2017). The gnostron framework pivots on the notion that mechanisms of neuronal groupings envisioned in the VAN map directly onto heuristics for probabilistic resource optimization so that energy savings in the biological substrate equate to reduced processing expenditures in the machine implementation. **Figure 13** returns to hunting frogs (section "Integrating Virtual Associative Network into the Variational Free Energy Minimization Framework") in order to illustrate and summarizes these notions,

Boundary energy barriers bound evidence for the corresponding object (Yufik and Friston, 2016; Yufik, 2019, 2021a,b). More precisely, recognition confidence associated with firing a neuron is a function of the corresponding probability μ_{ih} in the neuron's receptive field and the strength of neuron's attachment to (correlation with) other neurons in the packet. High confidence motivates leaving the packet but the fee charged for crossing the barrier discourages premature decisions and forces seeking confirmation or disconfirmation, in which case paying the fee remains the only option.

Technically, formation of packets constitutes a heuristic yielding complexity reduction in the probabilistic optimization problem. More precisely, forming packet network atop the associative network breaks a very large, continuous problem into a succession of discrete problems small enough to be solved by full search (this strategy appears to underlie the Long Term Memory/Short Term Memory (STM) architecture where small STM buffer [less than 10 items (Miller, 1956)] is subject to exhaustive scanning (Sternberg, 1969). The computational architecture of associative cortices readily affords self-partitioning in associative networks allowing near-optimal behavior. In Gnostron, the partitioning quality is defined by a simple criterion: choose a particular optimization algorithm and compare results obtained before (baseline) and after partitioning into packets [a stripped down, proof-of-concept system for target recognition obtained close to two orders of magnitude complexity reduction with acceptably small error amplitude (Malhotra and Yufik, 1999; Yufik and Malhotra, 1999)]. **Figure 13** generalizes the gnostron proposal.

Figure 14 lists key neuronal mechanisms postulated in VAN, seeking to establish three points: First, the postulated mechanisms have algorithmic expression in the framework of probabilistic resources optimization. Second, gnostron framework establishes a degree of isomorphism between human decision processes (as envisioned in VAN) and computational procedures: both substantive decision-making and Gnostron procedures operate with models representing objects, behavior and relations. Moreover, lower level gnostron procedures can be mapped meaningfully onto mental operations (for example, computing packets involves operations on cutsets in networks that correspond, roughly, to refocusing attention from prominent relations between objects to background relations that were deemed to be less significant). Finally, gnostron mediates between human operators and other systems but does not replace them (for example, gnostron can be calling on standard on-board systems to estimate the chances of safe landing in the New Jersey Airport. By the same token, it will be able to respond to a query like "Is the NJA an option?"). In this way, gnostron shields an operator from computation details while maintaining interaction at a substantive level adequate for shared situational understanding.

Strategy V involves formation of fixed templates. To appreciate differences in performance yielded by strategies V and VI, map them onto acquisition of chess skills, as follows: strategy III enables one to tell apart (recognize) chess pieces, strategy IV associates admissible behavior (rules) with the pieces, and strategy V enables memorization of particular tactics. To take a

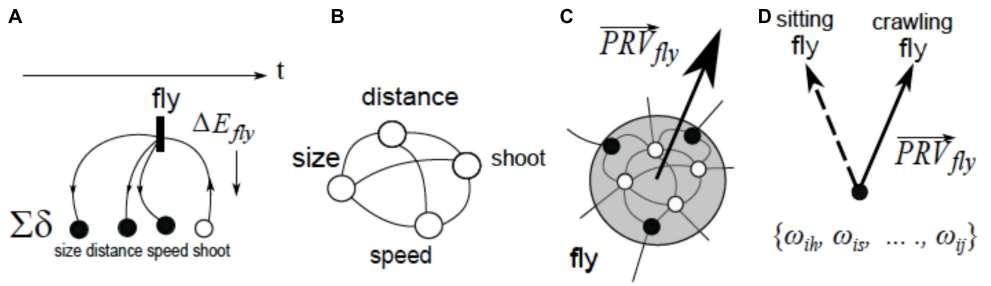


FIGURE 13 | Frogs shoot their tongues at small moving objects in sufficient proximity. **(A)** Mobilizing, firing and inhibiting neurons consumes energy and extracts energy deposits from the world stream (the flies). **(B)** Connections are formed between co-firing neurons and get strengthened with each successful episode (these connections can be genetically fixed, as in frogs). **(C)** Phase transition in the network turns the associative group into a cohesive packet bounded by an energy barrier. **(D)** The packet is a functional unit (has a receptive field computed as a function of receptive fields in the constituent neurons) amenable to mobilization and allocation. In the humans, mobilizing and allocating packets is experienced as perceiving "objects," packet "tuning" (rotating packet vector) defines the states of the object, admissible transitions between the states define behaviors (self-directed packet manipulation is available to humans but not to frogs).

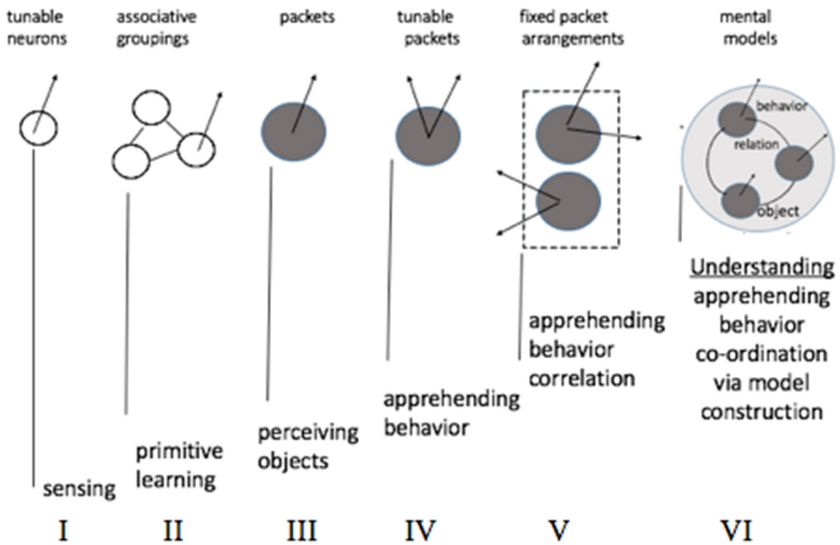


FIGURE 14 | VAN identifies six basic mechanisms employed by the brain to produce adaptive responses to variations in the world stream. The mechanisms vary from tuning receptive fields in individual neurons **(I)** to constructing and manipulating mental models **(VI)**. Gnostron architecture integrates these mechanisms.

closer look at the latter, a few chess notions will be helpful: Fool's mate (capital F) is a checkmate delivered in the fewest possible moves (2–4) after the beginning of the game, fool's mate (small f) is a maneuver of a few moves anytime in the game that delivers checkmate or turns opponents' position into a hopelessly lost one, and Sicilian defense is a particular Black move in response to a particular White move at the opening of the game (1. e4 c5). It's easy to see that a novice player taught only the Sicilian template is unlikely to seek tournaments ("what will happen after I do c5?"). Chess books teach seven basic strategies for continuing the game but, being taught all seven or, to take things to the extreme, having memorized the gazillion games ever played that used Sicilian template would make no difference: fool's mate is guaranteed if a more skilled opponent deviates from one of the memorized games, or just opens the game by any move other than e4.

The argument is (a re-statement of Searle's Chinese room argument) that knowledge, however, extensive, neither amounts nor guarantees understanding. Moreover, knowledge without understanding easily becomes a vulnerability. More to the point, the German army delivered fool's mate to the French command at the beginning of the campaign, taking advantage of the fact that the latter adhered to a rigid tactical template acquired in the WWI. Deficiencies in strategic thinking on that scale can hardly be remedied by a decision aid (although detecting rigid templates can be a part of gnostron tactics when interacting with human operators).

The chess example will serve to illustrate a general contention regarding situational understanding in both the human and the machine, as follows. Understanding involves the ability to form templates that is inextricably combined with the ability to re-structure and deviate from them and to incorporate them as units into other structures. Growing understanding is accompanied by growing organization and global order in the neuronal pool (see **Figure 13**) and growing repertoire of sensorimotor activities (e.g.,

from acquiring a repertoire of standard procedures in managing routine flights to safely ditching a suddenly disabled aircraft). The expanding activity repertoire entails growing entropy in the sensory-motor system. That is, understanding capacity brings about reduction of entropy in the internal states while increasing entropy in the motor-sensory periphery. **Figure 15** illustrates this important aspect of VAN/VFEM integration.

Incorporation of understanding into the VFEM schema, as in **Figure 14**, suggests a modification in the formulation of the principle, as follows: $F_V \rightarrow$ min under $H_{MB} \rightarrow$ max and $W(D_{KL}) \rightarrow$ max.

Here, $W(D_{KL})$ denotes the amount of work invested in minimizing discrepancy between the predicted and actual probability distributions [the Kullback-Leibler divergence was shown to define a lower bound to entropy production and thus the average amount of work dissipated along the process (divided by the temperature) (Roldan and Parrondo, 2012)].

That is, under a fixed energy budget in the brain, understanding capacity is a result of increased organization (decreased entropy) in the regulatory system which diverts more energy to—and thus increasing the amount of useful work in—the memory system, to allow expanding activity repertoires (growing entropy) in the motor-sensory system (see **Figure 2**) that in turn leads to increasing (and/or stabilizing) energy inflows extracted from the world stream. Operations on models underlie prediction and retrodiction: in A and B under relation r, changes in the behavior of A predict changes in the behavior of B and changes in the behavior B retrodict to changes in the behavior in A, as afforded by the relation r. The process is tightly constrained in a template (e.g., under conviction that frontal assault is the only viable strategy, any intelligence is interpreted as either conforming, or irrelevant, or a product of deliberate misinformation). Transition from template-matching to mental modeling relaxes the constraints, posing the problem of hypotheses selection ("that does not look like preparations for a frontal assault, what can that possibly be?") captured in the notion of abductive inference.

> "The first starting of a hypothesis and the entertaining of it, either as simple interrogation or with any degree of confidence, is an inferential step which I propose to call abduction (or retrodiction). This will include a preference for any hypothesis over others which would equally explain the facts, so long as as this preference is not based on upon any previous knowledge bearing upon the truth of the hypotheses, nor on any testing of any of the hypotheses, after having admitted them on probation. . . . the whole question of what one of the number of possible hypotheses ought to be entertained becomes purely a question of economy " (Peirce, 1901/1955, pp. 151, 154).

The thinking process naturally selects the path of least resistance (i.e., strong associations, as in a template), and needs to be forcefully interrupted and re-directed to paths deviating from "any previous knowledge." These operations are defined as "intervention" and insertion of "counterfactuals" in a recent probabilistic model of causal reasoning (Pearl and Mackenzie, 2018), and are represented by operations of jump, coordination and blending in VAN (to be discussed elsewhere).

To summarize, this section mapped some of the cognitive operations claimed to underlie understanding capacity in the humans onto computational procedures defined within the probabilistic optimization framework [excepting some residue having no computational expression (Penrose, 1997)]. It was proposed that understanding allows the brain to deal with non-contiguous, weakly correlated stimuli groupings in the world stream. In particular, understanding makes possible accounting for complex interdependencies between actions and world states, as in producing changes in objects indirectly, *via* coordinated changes in some other objects. Cognitive operations boil down to variable grouping and stabilization of the groups which enables subsequent intra-group variation and inter-group coordination, all serving to maximize and stabilize energy rewards (value) while minimizing internal energy costs. These operations can be mapped onto brain components whose functions have been defined in classical models as well as in some recent findings [e.g., the hippocampus has been found to be constructing abstract values spaces (Knudsen and Wallis, 2021)]. Emphasizing the role of coordination in understanding is consistent with a classical theory (Piaget, 1975, 1978) and with some recent findings concerning the role of cerebellum in the higher cognitive functions (Cerminara et al., 2009; Schmahmann et al., 2019).

Implementing operations postulated in the cognitive theory in tractable algorithms would endow machines with capabilities approximating those attributed to human situational understanding within selected situation classes. In particular, machines could be approaching the ability to "feel" the direction of appropriate actions without examining details and to formulate recommendations, explain them and receive instructions from human operators expressed in substantive terms. Attaining situational understanding reduces operational complexity (Yufik and Hartzell, 1989) enables explainable predictions, identification of critical situational elements and dynamic orchestration and optimization of cognitive and computing resources (Lieder and Griffiths, 2019).

DISCUSSION

Arguably, foundational ideas of the cognitivist framework were influenced by von Neumann's conceptualization of computing systems envisioning that data and procedures for operating on the data are held in the same medium. The template "data – procedures" holds no "slots" for understanding so adopting the template in representing cognition required marginalizing the role of that capacity in intelligent performance. Accordingly, a definitive volume on human problem-solving mentioned understanding once in the concluding chapters, and only to point out that "high level of mechanization can be achieved in executing the algorithm, without any evidence of understanding" (Newell and Simon, 1972, p. 832). The cognitivist framework accorded understanding no function in the architecture of cognition (Anderson, 1983; Rosenbloom et al., 1991] nor any place in a theory of cognition (Newell, 1992), and structured the definition of understanding so it could be forced into the available two "slots":

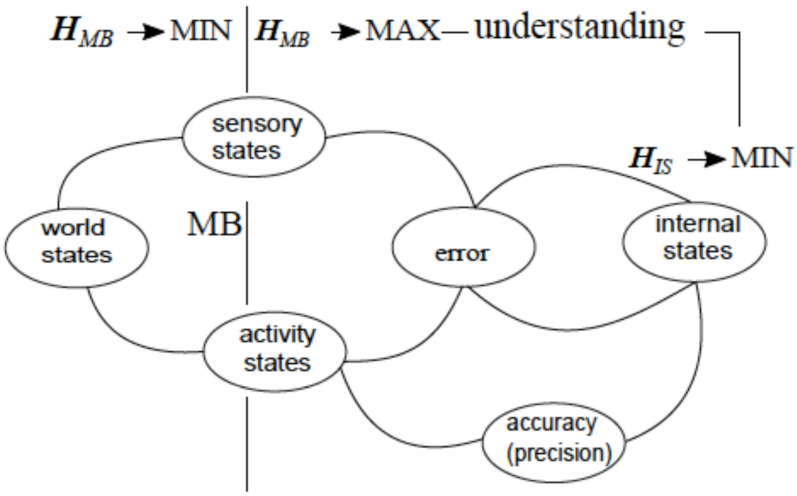


FIGURE 15 | An absence of understanding capacity entails tendency to minimize the entropy in a Markov Blanket while understanding seeks to maximize entropy in MB while minimizing entropy of the internal states [entropy of associative network is maximal if potential connectivity is unrestricted (**Figure 14I**) and minimal when connections are restricted (by coordination demands) and sparse (**Figure 14VI**)].

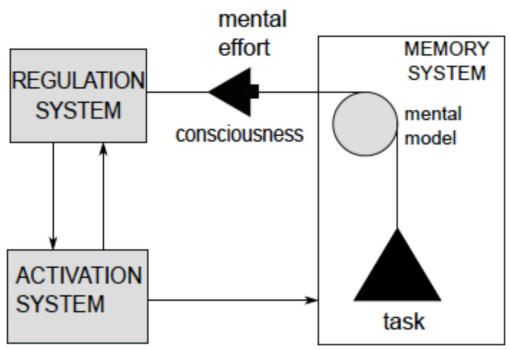


FIGURE 16 | Adapting to world streams involves work performed by regulatory system on a memory system. Mental models are neuronal structures amplifying the brain's capacity to handle challenging tasks (non-contiguous stimuli, acting on objects to impact other objects, etc.) within a limited energy budget (Raichle and Gusnard, 2002) (question "who is pulling the rope in the pulley?" will be entertained elsewhere).

> "S understands knowledge K if S uses K whenever appropriate. S understands task T if S has knowledge and procedures needed to perform T" (Simon, 1979, p. 447).

Language understanding was conceptualized as manipulation of scripts (i.e., template matching) (Schank and Abelson, 1977). It is interesting to note that a book addressing the practice of problem solving as opposed to the theory of that in Newell and Simon (1972), presented in the front-page picture some key notions that were overlooked in the theory: the brain was depicted as a contraption comprising a power plant, a regulator and a system of wheels delivering power to a pulley used for lifting weights (Fogler et al., 2013). **Figure 16** borrows from that depiction to re-state a main message of this paper.

Conceptualizing cognition as mental work invested in dynamic orchestration and marshaling of neuronal resources suggests a simple definition of consciousness, as follows (we are taking the liberty of citing an earlier work):

> "Virtual networks form spontaneously on top of the associative network. By contrast, operations on the virtual network are not spontaneous but self-directed (deliberate, attentive, conscious) and are conducted by the control module. These operations perform work and require cognitive effort, the term "consciousness" denotes the experience of exerting that effort. On that account, "cogito ergo sum" expresses not an inference but a direct experience of cognitive strain – one can doubt the reality of the objects of thinking and even of the subject of thinking but cannot doubt the immediate and direct experience of an effort exerted in the process of thinking" (Yufik, 2013, p. 50).

In short, VAN suggested that cognitive processes alternate between conscious (deliberate, effortful) and subconscious (spontaneous) phases. It is encouraging that later studies have arrived at similar conclusion in treating the phenomenon of consciousness (Solms, 2021).

In general, the VAN approach allowed drawing a line from neuronal processes all the way up to understanding and consciousness. The line is admittedly thin and punctuated but short (only 4 waypoints), connecting basic experimental findings ["tunable" neurons (Fritz et al., 2003, 2005), "tunable" assemblies (Georgopoulos and Massey, 1987; Georgopoulos et al., 1989, 1993)] to most advanced cognitive theories (Friston and Stephan, 2007; Friston, 2009, 2010; Parr and Friston, 2019; Ramstead et al., 2021). The approach builds on some of the key insights at the foundation of cognitive science [neuronal assemblies (Hebb, 1949, 1980), fluid intelligence (Cattell, 1971, 1978), mental effort in memory retrieval (James, 1950/1890), understanding as co-instantaneous co-ordination (Piaget, 1978, 1975), mental modeling (Johnson-Laird, 1983), other], and anticipated some

of the recent ideas and suggestions relating energy processes and cognition (Christie and Schrater, 2015; Pepperell, 2018; Vergara et al., 2019; Hylton, 2020). Consistent with the recent analysis of different kinds of free energy in the Bayesian account of cognition (Gottwald and Braun, 2020), the VAN model establishes reciprocity between the minimization of variational free energy and minimization of thermodynamic free energy in the neuronal system (Yufik and Friston, 2016; Yufik et al., 2016). Tentatively, the approach suggested unity of or a close relation between the mechanisms of sensori-motor coordination (Sparrow and Irizarry-Lopez, 1987; Sparrow and Newell, 1998; Sparrow et al., 2007; Latash, 2008, 2021), cortical coordination (Bressler and Kelso, 2001) and coordination in mental models (Yufik and Friston, 2016; Yufik, 2019, 2021a,b). Finally, the approach informs design of operator support in complex dynamics tasks (Yufik and Hartzell, 1989; Yufik and Sheridan, 1997; Landry et al., 2001; Yufik and Sheridan, 2002) using a transparent mathematical formalism (Yufik, 1998). Arguably, the hierarchy of VAN processing mechanisms (as in **Figure 13**) is compatible with the idea of "neuron-centered concepts" that associates concepts with patterns of input information evoking specific selective responses in groups of neurons (Gorban et al., 2019) [VAN postulates existence of complex neurons responding to specific activity patterns in lower-level (simpler) neurons or neuronal groupings]. The VAN view is consistent with the notion of cognition grounded in modal simulations, bodily states, and situated actions (Barsalou, 2008), as opposed to more conventional view in AI reducing cognition to computations on amodal symbols. As was argued earlier in this section, the conventional (cognitivist) approach has been downplaying the role of understanding in intelligent performance.

With some exaggeration, the view on cognition adopted in AI and cognitive science can be characterized as "intelligence without understanding." Figuratively, human intelligence can be compared to an Egyptian pyramid visited by tourists who are paying attention to a few stones at the bottom (learning) and the last stone on top (reasoning) while ignoring the rest. The pyramid holds a great promise since even limited explorations have produced spectacular successes. In the period of about 60 years, during which neural network technology has progressed from handling simple tasks (like recognizing letters) to participating in the most complex form of scientific analysis (Krasnopolsky, 2013) and beating humans in the games of chess and Go. The technology is based on algebraic methods of iterative error reduction (training) which are highly computationally intense. Accordingly, the progress was due to increases in hardware efficiency and the development of ingenious heuristics aimed at reducing the computational complexity of the iteration procedures. The hardware efficiency has increased about a billion times [NVIDIA's GTX 1080 GPU delivers nine teraflops for about $ 500, a similar power output in 1961 would have cost about $9 trillion for a string of IBM 1620 computers (Shepard et al., 2018)]. For argument's sake, assume that the efficiency of the procedures has increased a thousand times, yielding a trillion times increase in the overall efficiency. Consider the following: the analysis of eye movements showed that expert chess players immediately and exclusively focused on the relevant aspects in the chess task while novices also examined irrelevant aspects (Bilalić et al., 2010). The ability to "feel" the situation, or to "know what should happen in given circumstances" prior to examining those circumstances in detail [(Feynman, c/f de Regt, 2017, p. 102] makes possible competition between slow thinking human players and fast computing chess machines.

The point is that the brain cannot accelerate either the underlying biophysical processes or the conscious reasoning, can neither miniaturize neurons nor increase their number, and cannot significantly increase the average rate of ATP production. These limitations foreclosed the paths to cognitive performance improvements taken in AI and enforced development of radically different strategies. A fair competition between human players and chess algorithms would require running the algorithms on an abacus or some manual calculator.

AI is being widely perceived as a critical and, perhaps, decisive component in the national defense (West and Allen, 2020; Niotto, 2021), giving an advantage that derives predominantly from the strength of machine learning in general and neural nets in particular. The expectation seems to be that friendly neural nets will be victorious over the adversarial ones, which calls for designing methods to deceiving adversarial nets (e.g., Nguyen et al., 2015) while ruggedizing own nets and preparing them for frontal assaults. Conclusion of an expert group tasked with assessing the implementation of AI for the Department of Defense appear to be curbing the expectation:

> "the sheer magnitude, millions of billions of parameters (or weights) which are learned as part of the training... makes it impossible to really understand exactly how the machine does what it does. Thus the response of the network to all possible inputs is unknowable" (Scharre, 2018, p. 186).

It is interesting to note that recent developments in the neural net technology have taken a turn suggesting possible convergence with some of the methods outlined in this paper. In particular, clusters of neurons (called "capsules") are being identified in neural nets whose activity vector is taken to constitute the instantiation parameters of a specific type of entity such as an object or an object part. With that, the length of the activity vector is taken to represent the probability that the entity exists and its orientation to represent the instantiation parameters (Sabour et al., 2017).

To main points in this paper can be summarized as follows:

1. The paper presented a definition of understanding that is consistent with and substantiating analysis of understanding capacity in the current literature (Piaget, 1978; de Regt, 2017), outlined several hypotheses concerning the underlying mechanisms (the VAN theory) and suggested that (a) understanding constitutes a special form of Active Inference and (b) situational understanding enables situation awareness, consistent with the conceptualization expressed in **Figure 1**.
2. The active inference framework encompasses the entire spectrum of living organisms and associates adaptive

behavior with the minimization of variational free energy in the nervous system (Friston, 2010). According to VAN, understanding engages mechanisms that are unique to humans and yield a dual benefit of decreasing both the variational free energy and the metabolic energy expenditures. Minimization of variational free energy roughly equates to minimizing prediction error. Prediction *via* understanding provides a uniquely efficient form of error reduction.

3. The notion that minimization of metabolic costs can serve as a unifying principle in considering brain processes is not new (e.g., Hasenstaub et al., 2010; Huang et al., 2012). The VAN proposal deviates from the other suggestions, by (a) identifying specific mechanisms of metabolic cost minimization and (b) associating these mechanisms with a potentially unlimited growth in the variety and complexity of tasks accessible to humans, including the ability to overcome the inertia of past learning and to act efficiently under fluid and novel circumstances having no past precedents (Yufik and Sheridan, 2002; Yufik, 2013).
4. Understanding involves self-directed composition of coordinated neuronal structures (mental models) establishing relations (dependencies) between entities perceived previously as separate and independent. Composing such models can be highly effort-demanding. However, such composition expenditures are compensated by low-effort manipulations of the models making one aware of how local changes can bring about and coordinate with changes in the rest of the model (e.g., Yufik and Yufik, 2018). More precisely, manipulating models can "give some feel for the character" of coordinated changes (de Regt, 2017), which subsequently focuses attention on the critical situation elements. In general, mental modeling enables advances in the performance of complex tasks, by minimizing both the internal costs of the foresight and the risk of costly errors.
5. The paper used the notion of binary neurons, but only to simplify the argument. The theory is not restricted to this simplification, hypothesizing the existence of classes of complex neurons responding to different activity patterns in their input, to combinations of such activity patterns in several neuronal groups, or to forms of activity coordination [e.g., "concept cells" responding to different images of a person as well as the written and spoken names of that person (Quiroga, 2020) belong to the second class]. The pivotal notion of packets defines a property of neuronal groups that is invariant across models of neurons [in the same way as the notion of "neuronal assembly" (Hebb, 1980) is not committed to any particular model]. The theory builds on two experimentally established and model-invariant characteristics of neuronal mechanisms [rotation of assembly vectors (Georgopoulos et al., 1989) and task-related plasticity of neuronal receptive fields (Fritz et al., 2003)], expanding their application to complex neurons and neuronal groupings.
6. The theory derives understanding from coordination in the behavior (patterns of excitation-inhibition activities) of neuronal packets, which is consistent with conceptualizing brain as a dynamical system or "dynome" [as opposed to static "connectome" (Kopell et al., 2014)]. By definition, virtual network comprises a hierarchy of network types [synaptic, associative, packet, behavioral and relational networks (Yufik, 1998, 2019)]. Roughly, the former two network types belong to neural and functional connectomes while the latter three types form a dynome. Recent literature associates advanced cognitive capabilities in primates and humans with the ability to monitor the significance of multiple goals in parallel, and to switch between the goals (Mansouri et al., 2009; Mansouri et al., 2017). The present proposal expands the scope of advanced capabilities in the humans, to include dynamic coordination of multiple goals within integrated situation models.
7. The paper argues that increasing the efficiency of human-machine systems, particularly in challenging circumstances (short decision cycle, high cost of errors, etc.) requires mutual understanding between the parties. The VAN theory suggests an avenue toward meeting the requirement, offering tractable procedures amenable to integration with the methods of active inference. The VAN formalism (gnostron) is orthogonal to methods rooted in the perceptron architecture (vector movement coordination in dynamically composed networks in the gnostron vs. vector mapping in fixed networks with adjustable synaptic weights in the perceptron).
8. Mental modeling constitutes a form of self-organization in the brain. Biological processes underlying such self-organization can be approximated computationally in conventional (von Neumann-Turing) computers or, potentially, emulated in devices operating on principles different from those adopted in the conventional machines (Hylton, 2020).
9. In machine understanding, as conceptualized in VAN, machine processes and human cognitive processes are isomorphic, i.e., humans think of entities, behavior and relations and machines compute the same. Shared situational understanding in a human-machine system does not make the system infallible but can be expected to amplify and accelerate human grasp, increase human trust and confidence, and sharply reduce the likelihood of costly errors. In the autonomous scenarios, understanding expands the range of tasks that can be reliably delegated to the machine (methods for measuring performance improvements resulting from machine understanding are beyond the scope of this paper).

The above points suggest directions for further R&D, from developing deeper insights into the role and mechanisms of understanding to formulating tractable computational formalisms and designing artifacts that take advantage of those insights. The VAN/VFEM proposal contends that the objective of ensuring battlespace dominance brings to the fore the problem of situation understanding enabling coordination and prediction of

multiple activities under conditions that might be unfamiliar and undergoing kaleidoscopic changes. The proposal complements advances in machine learning and suggests other approaches that might be worth exploring.

It feels appropriate to conclude the discussion with a quote from a philosopher of mind and Nobel Laureate in physics:

> ". . .it seems to me that intelligence is something which requires understanding. To use the term intelligence in a context in which we deny that any understanding is present seems to me unreasonable. Likewise, understanding without any awareness is
>
> also a bit of a non-sense. . . . So that means that intelligence requires awareness. Although I am not defining any of these terms, it seems to me to be reasonable to insist upon these relations between them" (Penrose et al., 2000, p. 100).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YY contributed theoretical discussion and wrote the manuscript. RM contributed to the discussion and analyzed examples and applications.

FUNDING

YY received funding from Virtual Structures Research Inc, a non-profit U.S. company.

ACKNOWLEDGMENTS

RM is Senior Research Engineer in the United States Air Force Research Laboratory Sensors Directorate, WPAFB, OH, United States. The authors are indebted to Karl Friston for detailed feedback and insightful suggestions. The authors also thank reviewers for critique and comments that helped improving the manuscript. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

REFERENCES

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Avery, M. C., and Krichmar, J. L. (2017). Neuromodulatory systems and their interactions: a review of models, theories, and experiments. *Front. Neural Circuits* 11:108. doi: 10.3389/fncir.2017.00108

Barsalou, L. W. (2008). Grounded cognition. *Ann. Rev. Psychol.* 59, 617–645.

Berry, J., Brangwynne, C. P., and Haataja, M. (2018). Physical principles of intracellular organization via active and passive phase transitions. *Rep. Prog. Phys.* 81:046601. doi: 10.1088/1361-6633/aaa61e

Berwick, R. C., and Chomsky, N. (2016). *Why Only Us? Language and Evolution*. Cambridge, MA: The MIT Press.

Bilalić, M., Langner, R., Erb, M., and Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: a study with chess experts. *J. Exp. Psychol. Gen.* 139, 728–742. doi: 10.1037/a0020756

Bressler, S. L., and Kelso, J. A. (2001). Cortical coordination dynamics and cognition. *Trends Cogn. Neurosci.* 5, 26–36.

Cattell, R. B. (1971). *Abilities: Their Structure, Growth, and Action*. New York, NY: Houghton Mifflin.

Cattell, R. B. (1978). *Intelligence: Its Structure, Growth and Action*. New York, NY: Elsevier.

Cerminara, N. L., Koutsikou, S., Bridget, M., Lumb, B. M., and Apps, R. (2009). The periaqueductal grey modulates sensory input to the cerebellum: a role in coping behaviour? *Eur. J. Neurosci.* 29, 2197–2206. doi: 10.1111/j.1460-9568.2009.06760.x

Chomsky, N. (2007). Biolinguistic explorations: design, development, evolution. *Int. J. Philos. Stud.* 15, 1–21.

Christie, S. T., and Schrater, P. (2015). Cognitive cost as dynamic allocation of energetic resources. *Front. Neurosci.* 9:289. doi: 10.3389/fnins.2015.00289

Clausewitz, C. (2015/1835). *On War*. Scotts Valley, CA: CreateSpace Independent Publishing.

de Regt, H. W. (2017). *Understanding Scientific Understanding*. New York, NY: Oxford University Press.

Endsley, M. R. (1987). *SAGAT: A Methodology for the Measurement of Situation Awareness*. Northrop Technical Report NOR DOC 87-83. Hawthorne, CA: Northrop Corp.

Endsley, M. R. (1988). "Design and evaluation for situation awareness enhancement," in *Proceedings of the Human Factors Society: 32nd Annual Meeting*, Anaheim, CA, 97–101. doi: 10.1177/154193128803200221

Endsley, M. R. (1994). "Situation awareness in dynamic human decision making: theory," in *Situational Awareness in Complex Systems*, eds R. D. Gilson, D. J. Garland, and J. M. Koonce (Daytona Beach, FL: Embry-Riddle Aeronautical University Press), 27–58.

Endsley, M. R., and Connors, E. S. (2014). "Foundations and challenges," in *Cyber Defense and Situational Awareness*, eds A. Kott, C. Wang, and R. E. Erbacher (New York, NY: Springer), 7–29.

Fogler, H. S., LeBlanc, S. E., and Rizzi, B. R. (2013). *Strategies for Creative Problem Solving*. New York, NY: Pearson.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Fritz, J. B., Elhilali, M., and Shamma, S. A. (2005). Active listening: task-dependent plasticity of receptive fields in primary auditory cortex. *Hear. Res.* 206, 159–176. doi: 10.1016/j.heares.2005.01.015

Fritz, J. B., Shamma, S. A., Elhilali, M., and Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Georgopoulos, A. P., Kettner, R. E., and Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* 1988, 2928–2937. doi: 10.1523/JNEUROSCI.08-08-02928.1988

Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., and Massey, J. T. (1989). Mental rotation of the neuronal population vector. *Science* 243, 234–236. doi: 10.1126/science.2911737

Georgopoulos, A. P., and Massey, J. T. (1987). Cognitive spatial-motor processes 1. The making of movements at various angles from a stimulus direction. *Exp. Brain Res.* 65, 361–370. doi: 10.1007/BF00236309

Georgopoulos, A. P., Taira, M., and Lukashin, A. (1993). Cognitive neurophysiology of the motor cortex. *Science* 1993, 47–52. doi: 10.1126/science.8465199

Gorban, A. N., Makarov, V. A., and Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Phys. Life Rev.* 29, 55–88.

Gottwald, S., and Braun, D. A. (2020). The two kinds of free energy and the Bayesian revolution. *PLoS Comput. Biol.* 16:e1008420. doi: 10.1371/journal.pcbi.1008420

Gu, S., Betzel, R. F., Mattar, M. G., Ciesiak, M., Delio, P. R., Graffon, S. T., et al. (2017). Optimal trajectories of brain state transitions. *Neuroimage* 148, 305–317.

Gu, S., Cieslak, M., and Baird, B. (2018). The energy landscape of neurophysiological activity Implicit in brain network structure. *Sci. Rep.* 8:2507. doi: 10.1038/s41598-018-20123-8

Hasenstaub, A., Otte, S., Callaway, E., and Sejnowski, T. (2010). Metabolic cost as a unifying principle governing neuronal biophysics. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12329–12334. doi: 10.1073/pnas.0914886107

Hebb, D. O. (1949). *The Organization of Behavior*. New York, NY: Wiley & Sons.

Hebb, D. O. (1980). *Essay on Mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

Huang, H. J., Kram, R., and Ahmed, A. A. (2012). Reduction of metabolic cost during motor learning of arm reaching dynamics. *J. Neurosci.* 32960, 2182–2190. doi: 10.1523/JNEUROSCI.4003-11.2012

Hylton, T. (2020). Thermodynamic neural network. *Entropy* 22:256. doi: 10.3390/e22030256

James, W. (1950/1890). *The Principles of Psychology*, Vol. 1–2. New York, NY: Dover Publications.

Johnson-Laird, P. N. (1983). *Mental Models*. Boston, MA: Harvard University Press.

Kang, J., Pae, C., and Park, H.-J. (2019). Graph-theoretical analysis for energy landscape reveals the organization of state transitions in the resting-state human cerebral cortex. *PLoS One* 14:e0222161. doi: 10.1371/journal.pone.0222161

Knudsen, E. B., and Wallis, J. D. (2021). Hippocampal neurons construct a map of an abstract value space. *Cell Press* 184, 4640–4650.e10. doi: 10.1016/j.cell.2021.07.010

Koehler, W. (1999). *The Mentality of Apes*. London: Routledge.

Kopell, N., Gritton, H. J., Whittington, M. A., and Kramer, N. A. (2014). Beyond the connectome: the dynome. *Neuron* 83, 1319–1328. doi: 10.1016/j.neuron.2014.08.016

Kozma, R., Puljic, M., Balister, P., and Bollobas, B. (2005). Phase transitions in the neuropercolation model of neural populations with mixed local and non-local interactions. *Biol. Cybern.* 92, 367–379. doi: 10.1007/s00422-005-0565-z

Krasnopolsky, V. M. (2013). *The Application of Neural Networks in the Earth System Sciences. Neural Networks Emulations for Complex Multidimensional Mapping*. New York, NY: Springer.

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cortex* 12, 477–485. doi: 10.1093/cercor/12.5.477

Landry, S. J., Sheridan, T. B., and Yufik, Y. M. (2001). Cognitive grouping in air traffic control. *IEEE Trans. Intell. Trans. Syst.* 2, 92–101.

Latash, M. L. (2008). *Synergy*. New York, NY: Oxford University Press.

Latash, M. L. (2021). Understanding and synergy: a single concept at different levels of analysis. *Front. Syst. Neurocsci.* 5:735406. doi: 10.3389/fnsys.2021.735406

Lieder, F., and Griffiths, T. L. (2019). Resource- rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43:e1. doi: 10.1017/S0140525X1900061X

Lin, L., Osan, R., and Tsien, J. Z. (2006). Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes. *Trends Neurosci.* 2006, 48–57. doi: 10.1016/j.tins.2005.11.004

Loftus, E., and Ketcham, K. (1996). *The Myth of Repressed Memory: False Memories and Allegations of Sexual Abuse*. New York, NY: St. Martin's Griffin.

Luria, A. R. (1973). *The Working Brain. An Introduction to Neuropsychology*. London: Penguin Books Ltd.

Luria, A. R. (1974). *Higher Cortical Functions in Man*. New York, NY: Basic Books.

Malhotra, R. P., and Yufik, Y. M. (1999). "Graph-theoretic networks for holistic information fusion," in *Proceedings of the IJCNN'99. International Joint Conference on Neural Networks*, Vol. 4, Washington, DC, 2796–2801. doi: 10.4155/fmc.12.128

Mansouri, F. A., Koechlin, E., Rosa, M. G., and Buckley, M. J. (2017). Managing competing goals—a key role for the frontopolar cortex. *Nat. Rev. Neurosci.* 18, 645–657.

Mansouri, F. A., Tanaka, K., and Buckley, M. J. (2009). Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. *Nat. Rev. Neurosci.* 10, 141–152.

Maurois, A. (1941). *Why France Fell*. London: John Lane / The Bodley Head Publishing.

Miller, G. A. (1956). The magic number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97.

Miller, N. L., and Shattuck, L. G. (2004). *A Process Model of Situated Cognition in Military Command and Control*. Monterey, CA: Naval Postgraduate School.

Nadim, F., and Bucher, D. (2014). Neuromodulation of neurons and synapses. *Curr. Opin. Neurobiol.* 29, 48–56. doi: 10.1016/j.conb.2014.05.003

Newell, A. (1992). Precis of unified theories of cognition. *Behav. Brain Sci.* 15, 425–492. doi: 10.1017/S0140525X00069478

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Hoboken, NJ: Prentice Hall.

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: high confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 427–436.

Niotto, M. (2021). *Artificial Intelligence and nuclear warfare. Is doomsday closer. Cyber Security and AI series*. Available online at: https://thesecuritydistillery.org/all-articles/artificial-intelligence-and-nuclear-warfare-is-doomsday-closer (accessed July 7).

Niven, J. E., and Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *J. Exp. Biol.* 211, 1792–1804. doi: 10.1242/jeb.017574

Norman, D. (1988). *The Psychology of Everyday Things*. New York, NY: Basic Books.

Oakley, D. A. (1981). Performance of decorticated rats in two-choice visual discrimination apparatus. *Behav. Brain Res.* 3, 55–69. doi: 10.1016/0166-4328(81)90028-0

Parr, T., and Friston, K. (2019). Generalized free energy and active inference. *Biol. Cybern.* 113, 495–513.

Pearl, J., and Mackenzie, D. (2018). *The Book of Why. The New Science of Cause and Effect*. London: Penguin Books.

Pedrosa, V., and Clopath, C. (2017). The role of neuromodulators in cortical plasticity. A computational perspective. *Front. Synaptic Neurosci.* 8:38. doi: 10.3389/fnsyn.2016.00038

Peirce, C. S. (1901/1955). "Abduction and induction," in *Philosophical Writings of Pierce*, ed. J. Buchler (New York, NY: Dover Publications).

Penrose, R. (1997). On understanding understanding. *Int. Stud. Philos. Sci.* 11, 7–20.

Penrose, R., Shimony, A., Cartwright, N., and Hawking, S. (2000). *The Large, the Small, and the Human Mind*. New York, NY: Cambridge University Press.

Pepperell, R. (2018). Consciousness as a physical process caused by the organization of energy in the brain. *Front. Psychol.* 9:2091. doi: 10.3389/fpsyg.2018.02091

Pepperell, R. (2020). Vision as an energy-driven process. *arXiv* [Preprint] arXiv:2008.00754,

Piaget, J. (1975). *The Development of Thought: Equilibration of Cognitive Structures*. New York, NY: The Viking Press.

Piaget, J. (1978). *Success and Understanding*. Cambridge, MA: Harvard University Press.

Quiroga, R. Q. (2020). Searching for the neural correlates of human intelligence. *Curr. Biol.* 30, R335–R338. doi: 10.1016/j.cub.2020.03.004

Raichle, M. E., and Gusnard, D. A. (2002). Appraising the brain's energy budget. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10237–10239. doi: 10.1073/pnas.172399499

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., and Friston, K. (2021). Multiscale integration: beyond internalism and externalism. *Synthese* 198, 41–70. doi: 10.1007/s11229-019-02115-x

Roldan, E., and Parrondo, J. M. R. (2012). Entropy production and Kullback-Leibler divergence between stationary trajectories of discrete systems. *arXiv* [Preprint] arXiv:1201.5613v1 [cond-mat.stat-mech], doi: 10.1103/PhysRevE.85.031129

Rosenbloom, P. S., Newell, A., and Laird, J. E. (1991). "Toward the knowledge level in Soar: the role of the architecture in the use of knowledge," in *Architectures for Intelligence*, ed. K. VanLehn (Mahwah, NJ: LEA).

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv* [Preprint] arXiv:1710.09829v2 [cs.CV], doi: 10.3390/cancers13194974

Schank, R., and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures*. Mahwah, NJ: Erlbaum.

Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War*. New York, NY: W.W. Norton & Co.

Schmahmann, J. D., Guell, X., Stoodley, C. J., and Halko, M. A. (2019). The theory and neuroscience of cerebellar cognition. *Annu. Rev. Neurosci.* 42, 337–364. doi: 10.1146/annurev-neuro-070918-050258

Shepard, L., Hunter, A., Karlén, R., and Balieiro, L. (2018). *Artificial Intelligence and National Security*. Washington DC: Center for Strategic International Studies.

Sigurdsson, T., and Duvarci, S. (2016). Hippocampal-prefrontal interactions in cognition, behavior and psychiatric disease. *Front. Syst. Neurosci.* 9:190. doi: 10.3389/fnsys.2015.00190

Simon, H. A. (1979). *Models of Thought*, Vol. 1–2. Haven, CT: Yale University Press.

Solms, M. (2021). *The Hidden Spring: Journey to the Source of Consciousness*. London: WW Norton and Company Ltd.

Solms, M., and Friston, K. (2018). How and why consciousness arises: some considerations from physics and physiology. *Conscious. Stud.* 25, 202–238.

Sparrow, W. A., and Irizarry-Lopez, V. M. (1987). Mechanical efficiency and metabolic cost as measures of learning a novel gross-motor task. *J. Mot. Behav.* 19, 240–264. doi: 10.1080/00222895.1987.10735410

Sparrow, W. A., Lay, B. S., and O'Dwyer, N. J. (2007). Metabolic and attentional energy costs of interlimb coordination. *J. Mot. Behav.* 39, 259–275. doi: 10.3200/JMBR.39.4.259-275

Sparrow, W. A., and Newell, K. (1998). Metabolic energy expenditure and the regulation of movement economy. *Psychon. Bull. Rev.* 5, 173–196.

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction time experiments. *Am. Sci.* 57, 421–457.

Suhir, E. (2012). Human in the loop: predictive likelihood of vehicular mission success and safety. *J. Aircr.* 49, 29–41. doi: 10.2514/1.c031418

Suhir, E. (2013). Miracle-on-the-Hudson: quantitative aftermath. *Int. J. Hum. Fact. Modell. Simulat.* 4, 35–63.

Suhir, E. (2018). *Human-in-The-Loop: Probabilistic Modeling of an Aerospace Mission Outcome*. Boca Raton, FL: Taylor & Francis.

Suhir, E. (2019). Short note – adequate trust, human-capacity-factor, probability-distribution-function of human non-failure and its entropy. *Int. J. Hum. Fact. Modell. Simulat.* 7, 75–83.

Suhir, E., Scataglini, S., and Paul, G. (2021). "Extraordinary automated driving situations: probabilistic analytical modeling of Human-Systems-Integration (HSI) and the role of trust," in *Advances in Simulation and Digital Human Modeling. AHFE 2020. Advances in Intelligent Systems and Computing*, Vol. 1206, eds D. Cassenti, S. Scataglini, S. Rajulu, and J. Wright (Cham: Springer), 323–329. doi: 10.1007/978-3-030-51064-0_41

Vergara, R. C., Jaramillo-Riveri, S., Luarte, A., Moënne-Loccoz, C., Fuentes, R., Couve, A., et al. (2019). The energy homeostasis principle: neuronal energy regulation drives local network dynamics generating behavior. *Front. Comput. Neurosci.* 13:49. doi: 10.3389/fncom.2019.00049

Watanabe, T., Masuda, N., Megumi, F., Kanai, R., and Rees, G. (2014). Energy landscape and dynamics of brain activity during human bistable perception. *Nat. Commun.* 5:4765. doi: 10.1038/ncomms5765

West, G. M., and Allen, J. R. (2020). *Turning Point: Policymaking in the Era of Artificial Intelligence*. Washington, DC: Brookings Institution Press.

Wiener, E. L., and Nagler, D. C. (1988). *Human Factors in Aviation*. New York, NY: Academic Press, Inc.

Yufik, Y. M. (1996). *Probabilistic Resource Allocation Allocation With Self-Adaptive Capabilities*. Available online at: https://patents.google.com/patent/US5586219

Yufik, Y. M. (1998). "Virtual associative networks," in *Brain and Values*, ed. K. Pribram (Mahwah, NJ: LEA Publishers), 109–177.

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M. (2018). "Gnostron: a framework for human-like machine understanding," in *IEEE Symposium Series Computational Intelligence SSCI 2018*, Bangalore, 136–145.

Yufik, Y. M. (2019). The understanding capacity and information dynamics in the human brain. *Entropy* 21, 1–38. doi: 10.3390/e21030308

Yufik, Y. M. (2021a). Laws of nature in action, perception and thinking. Comments on 'Laws of nature that define biological action and perception. *Phys. Life Rev.* 36, 9–11. doi: 10.1016/j.plrev.2020.12.003

Yufik, Y. M. (2021b). "Brain functional architecture and human understanding," in *Connectivity and Functional Specialization in the Brain*, eds T. Heinbockel and Y. Zhou (London: IntechOpen), 131–230.

Yufik, Y. M., and Friston, K. (2016). Life and understanding: origins of the understanding capacity in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Yufik, Y. M., and Hartzell, J. E. (1989). "Design for trainability: assessment of operational complexity," in *Designing and using human-computer interfaces*, eds J. Salvendy and L. G. Smith (Amsterdam: Elsevier).

Yufik, Y. M., and Malhotra, R. (1999). Information blending inn virtual associative networks: a new paradigm for sensor integration. *Int. J. Artif. Intell. Tools* 8, 275–290. doi: 10.1142/s0218213099000191

Yufik, Y. M., Sengupta, B., and Friston, K. (2016). Self-organization in the nervous system. *Front. Syst. Neurosci.* 11:69. doi: 10.3389/fnsys.2017.00069

Yufik, Y. M., and Sheridan, T. (1997). New framework for operator modeling and interface optimization in complex supervisory control systems. *Annu. Rev. Control* 179–195. doi: 10.1016/s1367-5788(97)00016-3

Yufik, Y. M., and Sheridan, T. (2002). Swiss army knife and Ockham's razor: modeling and facilitating operator's comprehension in complex dynamic tasks. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 32, 185–199.

Yufik, Y. M., and Yufik, T. (2018). "Situational understanding," in *Proceedings of the 7th International Conference on Advances in Computing, Communication and Information CCIT 2018*, Rome, 21–27.

Conflict of Interest: YY was employed by the company Virtual Structures Research, Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 06 January 2022
doi: 10.3389/fnsys.2021.782781

The Energy Homeostasis Principle: A Naturalistic Approach to Explain the Emergence of Behavior

Sergio Vicencio-Jimenez[1], *Mario Villalobos*[2], *Pedro E. Maldonado*[3] *and Rodrigo C. Vergara*[4]*

[1]The Center for Hearing and Balance, Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Escuela de Psicología y Filosofía, Universidad de Tarapacá, Arica, Chile, [3]Laboratorio de Neurosistemas, Departamento de Neurociencia & BNI, Facultad de Medicina, Universidad de Chile, Santiago, Chile, [4]Departamento de Kinesiología, Facultad de Artes y Educación Física, Universidad Metropolitana de las Ciencias de la Educación, Ñuñoa, Chile

It is still elusive to explain the emergence of behavior and understanding based on its neural mechanisms. One renowned proposal is the Free Energy Principle (FEP), which uses an information-theoretic framework derived from thermodynamic considerations to describe how behavior and understanding emerge. FEP starts from a whole-organism approach, based on mental states and phenomena, mapping them into the neuronal substrate. An alternative approach, the Energy Homeostasis Principle (EHP), initiates a similar explanatory effort but starts from single-neuron phenomena and builds up to whole-organism behavior and understanding. In this work, we further develop the EHP as a distinct but complementary vision to FEP and try to explain how behavior and understanding would emerge from the local requirements of the neurons. Based on EHP and a strict naturalist approach that sees living beings as physical and deterministic systems, we explain scenarios where learning would emerge without the need for volition or goals. Given these starting points, we state several considerations of how we see the nervous system, particularly the role of the function, purpose, and conception of goal-oriented behavior. We problematize these conceptions, giving an alternative teleology-free framework in which behavior and, ultimately, understanding would still emerge. We reinterpret neural processing by explaining basic learning scenarios up to simple anticipatory behavior. Finally, we end the article with an evolutionary perspective of how this non-goal-oriented behavior appeared. We acknowledge that our proposal, in its current form, is still far from explaining the emergence of understanding. Nonetheless, we set the ground for an alternative neuron-based framework to ultimately explain understanding.

Keywords: homeostasis, free energy principle, behavior, energy, neural network

Edited by:
Yan Mark Yufik, Virtual Structures Research Inc., United States

Reviewed by:
Todd L. Hylton, University of California, San Diego, United States
Vladimir Kranopolsky, National Oceanic and Atmospheric Administration (NOAA), United States

***Correspondence:**
Rodrigo C. Vergara
rodrigo.vergara_o@umce.cl

Received: *24 September 2021*
Accepted: *13 December 2021*
Published: *06 January 2022*

Citation:
Vicencio-Jimenez S, Villalobos M, Maldonado PE and Vergara RC (2022) The Energy Homeostasis Principle: A Naturalistic Approach to Explain the Emergence of Behavior. Front. Syst. Neurosci. 15:782781. doi: 10.3389/fnsys.2021.782781

INTRODUCTION

When an animal displays different behaviors, what are the primary processes occurring in the nervous system? How do neurons, neuronal networks, and ultimately the whole nervous system participate in behavior generation? This article argues that the nervous system unfolds autogenous mechanisms of energetic homeostasis, maintaining its energy equilibrium as a system.

In our view, the nervous system operates in the continuous present tense of its structural dynamics under strictly local rules of energy stability, without pursuing biological goals or adaptive adjustments for the organism. This spontaneous process of maintaining its energy balance occurs so that under statistically normal anatomical, physiological, and ecological conditions, it results precisely in those behaviors that prove to be adaptive for the animal.

This view of the nervous system corresponds, in essence, to what has been recently introduced as the Energy Homeostasis Principle (EHP; Vergara et al., 2019). This theoretical proposal draws strongly from the autopoietic theory of cognition in the sense of being strictly naturalistic (Maturana, 1978; Villalobos and Ward, 2015; Villalobos, 2015), and resonates, although with important nuances, with some aspects of the Free Energy Principle (FEP) approach in theoretical neuroscience (Friston and Stephan, 2007; Friston, 2010). The EHP does not hold that animal behavior and cognition arise only because the nervous system is a homeostatic energy system. If that were the case, we should observe cognition and complex behavior in any homeostatic energy system, as may occur in an open thermodynamic system that exhibits some degree of stability, such as tornadoes and stars (Ulanowicz and Hannon, 1987; McGregor and Virgo, 2011). Instead, the proposal is to realize that while we observe the behavior or the signs of cognition shown by an organism, its nervous system operates by simply following, in its own way, the EHP.

The nervous system is a homeostatic energy system, like other similar natural systems, but with significant structural and organizational features that make it unique. These features are essential because they explain why the nervous system, despite operating under the EHP, can generate phenomena such as animal behavior and cognition. The argument EHP asserts is that despite all the unique features we may find in the nervous system, it remains the fact that its operations follow, ultimately, homeostatic energy mechanisms.

This latter statement merits further discussion. When we speak of the unique features in the nervous system, we are not inviting the reader to picture mysterious non-natural features. All thermodynamic systems that maintain stability and integrity for the period they exist, long or short, have their own features related to their specific structural compositions and dynamic patterns. Candle flames and tornadoes are both dissipative structures that exhibit thermodynamic stability in their respective magnitudes or scales. However, only candle flames generate fast exothermic combustion reactions, radiate light, and illuminate a dark room. Conversely, tornadoes, not candle flames, can travel kilometers through large geographic areas, lifting and violently shaking heavy objects. There is nothing mysterious about these differences. They relate to each system's respective chemical and physical features, which must be considered to explain the varied phenomena associated with each system. What is a candle doing as a system when its flame radiates light and warms up our hands? From the systemic thermodynamic point of view, it is simply maintaining its stability and integrity as a dissipative system. When a tornado passes through a village and destroys the houses, what is it doing as a system? Again, from the systemic thermodynamic point of view, it is simply maintaining its stability and integrity as a dissipative system. But, if both systems are doing the same, how do they generate such different phenomena and results? The answer lies in the unique features of each system, the context in which they form, their material qualities, and so on.

The nervous system is a homeostatic energy system. Still, the specific way it manifests such quality given its biological (e.g., histological) composition, anatomical structure and physiological organization, its looped coupling with both the internal milieu and the external environment, its development within the organism, generate distinctive results and phenomena called behavior and cognition. In what follows, we will review the general systemic conditions that run for the nervous system.

GENERAL SYSTEMIC CONDITIONS

To understand the nervous system and the phenomena typically associated with its functioning (e.g., perception, motor control, language, and consciousness) it is crucial to examine its peculiarities and distinctive features as a system. However, it is equally important to consider the conditions that the nervous system shares with all natural systems, living and not-living, and according to which it must work. After all, what is fascinating about the nervous system is that, being a natural system (that is, a system that respects the laws, conditions, and principles that rule and restrict every natural system), it can generate phenomena as peculiar and exceptional as perceptual experience, understanding, consciousness, language, and intelligent reasoning.

This latter explanatory exercise is essential because, when facing extremely complex explanatory problems, it is usually tempting and easy to resort to the strategy of endowing the components and explanatory machinery of the system under study with the very special and complex properties we want to explain. For instance, this was the case with the explanation of the phenomenon of life. For an extended period, it was assumed that the components of living beings were unique in that they were endowed with a certain kind of vital force or energy that was not present in the components of inert objects (Bechtel and Richardson, 1998). We tried to explain life by postulating that the matter of which living beings are made was itself, somehow, living. Similarly, when facing the problem of explaining cognitive and mental phenomena, such as perception or intelligent reasoning, it is tempting to think of the nervous system, its components, and machinery, as if they themselves operated with protocognitive (subpersonal, automatic, unconscious) cognitive mechanisms, as if the nervous system was an epistemic agent dealing itself with alleged problems of uncertainty and lack of information, working on the base of hypotheses, inferences, predictions, error detection, and looking for evidence and hypothesis confirmation.

As the cases of biology and the problem of life teach us, the strategy of projecting the properties and capacities of the explanandum, even in a carefully sophisticated deflationary way,

into the explanatory substratum itself does not lead to adequate explanations. We think we would do better if we take the nervous system not as a cognitive agent but as a physical machine (Ashby, 1947) and try to understand its operation according to the conditions that rule every physical system in general. Doing this does not mean, of course, ignoring the particular features of the nervous system regarding its structure and organization; it just means understanding that such specific features do not set the nervous system apart from the rest of the natural systems.

Before we further develop our argument for a strict naturalistic approach to explain the emergence of behavior, we consider it essential to lay out some foundational concepts, so the reader can better consider the starting points. These points are not meant to provide an exhaustive characterization of the nervous system; far from that. However, combined, they should help us understand, in broad terms, the way the nervous system operates and generates some of the phenomena associated with its functioning. We consider the following premises:

1. The nervous system is non-teleological. Its dynamics are not driven by purposes or goals. As is the case with natural systems in general, the dynamics of the nervous system unfold following physical laws that are blind to purposes or goals (Villalobos and Ward, 2015).
2. The nervous system is non-normative. Its dynamics are not based on normative considerations such as what is (or might be) good or bad, adequate, or inadequate, beneficial, or harmful to the system itself or the organism. As is the case with natural systems in general, the dynamics of the nervous system unfold following physical laws that are blind to normative values (Villalobos and Ward, 2015).
3. The interactions of the nervous system with its surrounding systems, both intra- and extra-organism, are structural (i.e., physical, chemical, energetic) in nature, not epistemic, informative, or cognitive (Maturana, 2002). The nervous system is not an epistemic agent that collects and processes information, and its functioning is not oriented to knowing (inferring, predicting, guessing) anything (Villalobos, 2015).
4. The components of the nervous system, its neurons, and networks work through strictly local interactions, without "having in view" distal states, either intra- or extra-organism (Maturana and Varela, 1987).
5. The nervous system operates in its continuous structural present, without "having in view" non-current states, either past or future (Ashby, 1960; Maturana, 2008).
6. The nervous system, at the neuroscience scale of analysis, behaves deterministically (Ashby, 1960; Maturana, 1980). It is not a free agent that chooses, among a set of possibilities, what to do. The nervous system does what it does every instant because its structure at that instant simply allows no other action.
7. The nervous system is an open thermodynamic system that exchanges matter and energy with its surroundings.
8. The nervous system is a homeostatic system that, like all homeostatic systems, maintains certain stability and equilibrium in its physical parameters and shows the capacity to restore them when they are disturbed within specific ranges (Ashby, 1960).
9. Nervous systems, since their first formation in the embryonal stage, grow and develop in the continuous coupling, adaptation, and structural coherence with their biological surroundings and the extra-organism environment. This is a trivial condition for every system. Everything that begins to exist does so because the conditions for its emergence and existence are given. Every system emerges adapted to, or in structural coherence with, its surrounding conditions. This adaptation is conserved while the system exists as such and lost when the system ceases to exist.
10. A nervous system with normal anatomical and physiological development is always coupled in a loop with:
    (i) other physiological systems of the organism, such as the endocrine, immune, cardiovascular, and digestive systems.
    (ii) the external environment through specialized sensory organs and motor structures. Since these couplings are functionally closed as feedback loops, the nervous system always affects itself through them and thus maintains its homeostasis. At the same time, since these couplings arise in structural coherence and adaptation from the beginning (recall point 9), the self-centered homeostatic dynamics of the nervous system result in the conservation of the adaptation of the rest of the organism.
11. Complex enough nervous systems are hierarchically organized as second-order homeostatic systems, therefore exhibiting ultrastability and great flexibility (Ashby, 1960). Hierarchy, in this context, implies that some of the feedback loops of the nervous system (mentioned in point 10) operate at the first level of stability, whereas others operate over them at a higher level. In this functional organization, the higher level constraints but does not eliminate the degrees of freedom of the lower level, so the latter can deploy a considerable range of variability in its dynamics to the extent that does not disturb the equilibrium of the former. Because of this, from the point of view of the higher level of homeostasis, the lower level will appear to show not only adaptive or "useful" dynamics but also "neutral" or "useless" ones.

In the following sections, we will elaborate on the EHP considering this set of premises to produce a plausible explanation for behavior and, ultimately, understanding. We will start arguing how a naturalistic approach is required to disentangle proximate causes (cell operation) from distal causes (organism operations). Then, we will build over this conception to reinterpret neural processing without goal or purpose. We will also evaluate anticipatory behavior by means of the EHP and contrast it with the FEP. Finally, we will offer an evolutionary argument regarding how these apparently goal-directed-behaviors emerge from non-teleological mechanisms. Moreover, we will discuss how useless behavior may appear and may constitute a potential adaptive advantage in evolutionary terms.

SOME SPECIFIC CONSIDERATIONS ABOUT THE NERVOUS SYSTEM

One way to illustrate how neuronal interactions are restricted, and therefore, locally driven interactions dynamics, is to realize their context. When comparing the whole organism to its component cells, or even organs, it can be noted that cells are sensitive to completely different scales of physical phenomena (Southern et al., 2008; Dada and Mendes, 2011). For instance, swimming in a pool or the ocean makes little difference to an experienced swimmer, whereas doing so in an aqueous solution would be lethal to a cell (Pedersen et al., 2011). This becomes very clear at the spatial and temporal scales (Engel, 1980; Southern et al., 2008; Dada and Mendes, 2011; DiFrisco, 2017). For example, at the chemical level, cells are most sensitive to their direct environment, a space in the order of micrometers or smaller, whereas we, as organisms, are sensitive to phenomena in the order of millimeters and beyond. Regarding the time scale, the difference is equally remarkable. Most of our cells are replaced in our lifetime (DiFrisco, 2017), which means their time scale is significantly shorter than ours.

We may argue the specifics of these differences, such as up to what point the scales overlap, or how arbitrary it is even to state that such scales exist. However, the core of that observation goes beyond the scales themselves, the point being the phenomenological operational closure of a whole human being compared with a single cell is remarkably different. What I see as a hamburger is not the same experience for a cell. On the one hand, a cell is too tiny to perceive the hamburger as a whole, but also its potential interactions with it are different from those we would engage in. There is a difference between how we perceive and the actions we may perform given such perception; how we couple with objects in behavior. As such, even if we would acknowledge that a neuron or neural network could foresee something, it would be in a shorter time span and based on their local interactions.

The global concept of how local interactions build up hierarchically to behavior is depicted in **Figure 1A**, where we intend to remark local interactions. For instance, cells may interact directly with other cellular phenomena only. By doing so, they are structurally coupled with the environment, and if alive, maintain their energetic equilibrium and, therefore, their operational closure (close-loop arrow). Hierarchically, these local interactions may lead to population phenomena, such as synchronization. Given the intricate codependence between the actions of individual cells, a group of cells starts to behave as a unity, like a fish shoal showing coordinated movements (Herbert-Read, 2016), or eusocial insects, where survival is a matter of the colony and not only of the individual (Gillooly et al., 2010). In both examples, there are not unique individuals signaling what has to be done to the colony, but rather local interactions as one-to-one individuals produce these complex phenomena. For instance, the fish shoal seems to move like a wholly coordinated system, while this global property answers to individual interactions of one fish considering the movements of the fish right next to it (Herbert-Read et al., 2011). As such, complex systemic phenomena may occur driven by local interactions when sensorimotor actions of individual entities are codependent and intimately coupled (Bonabeau et al., 1997). This distinction is critical to avoid extrapolating system properties to local components; however, it raises some challenges. Given our aim to explain the emergence of behavior from a naturalistic viewpoint, the difference in sensitivity is challenging for at least two reasons. The first reason is the difficulty in establishing relationships between these levels; if they do not perceive the same phenomena, how are their dynamics aligned for survival? This complicates the development of causal explanations in biology. A similar situation was noticed 60 years ago by Ernst Mayr (1961) when he established that virtually all explanations of biological phenomena consisted of sets of proximate causes and sets of ultimate causes (or distal, given our framework). In Ernst Mayr's work, he illustrates the difficulties in establishing the causes of behavior, arguing that they can be attributed to the environment, physiology (including molecular mechanisms), or the interaction between the two. In this context, proximate causes would be those that control the organism's responses to immediate environmental factors (such as the sunrise regulation of the sleep-wake cycle in a mouse), while ultimate causes would be those that have an impact on the organism's survival (such as increased nocturnal activity in mice that decreases the probability of encountering predators). These ultimate causes are rooted in evolutionary mechanisms and have been incorporated into the system through generations of natural selection (Mayr, 1961). Therefore, under the EHP view, behavior emerges from the intersection of coupled local interactions, which keep cells alive, and evolutionary pressure, that permits local conditions to remain coupled, if they do not jeopardize the life of the whole organism (the distal cause). It is critical to notice that the distal cause can be interpreted as a consequence of meeting local requirements. Recalling point 6, "The nervous system does what it does at every instant because its structure at that instant simply allows no other action". In other words, distal causes exist as a result of living beings staying alive coupled with their environment and restricted to the evolutionary and individual history that has determined particular properties of their structure.

There is a second reason where local interactions are relevant. For the organism to survive, the fundamental needs of all these hierarchic levels must be met (**Figure 1A**). The specific needs of different kinds of cells are varied and different from those of the organisms they compose. Therefore, there are multiple layers or levels of operational closure that are not strictly equivalent nor overlapped and they must meet the entire organism requirements to stay alive and coupled with its environment. This illustrates the complex synchronization that must occur in the cell population of such an organism to survive, as well as the close codependence of a variety of cellular populations with remarkably different requirements.

Now, an apparent contradiction appears. Despite the short overlap of sensitivity to phenomena between the parts of our body and the whole organism, we exhibit adaptative behaviors. This supposed paradox has been solved mainly by assigning functions aimed at the survival of the entire organism to different parts of the body (Roux, 2014). However, this position usually

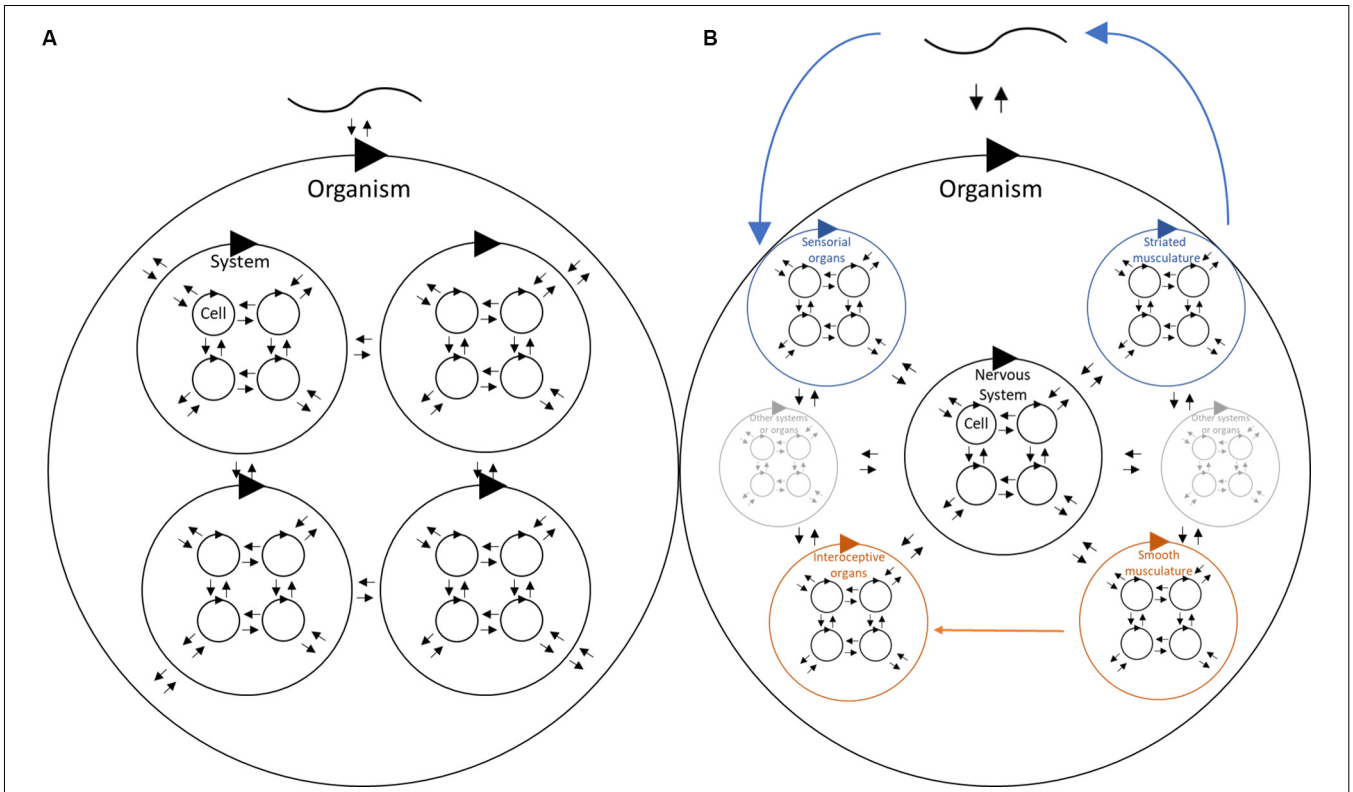


FIGURE 1 | Schematic hierarchical structure and interactions for **(A)** a generic living organism, and **(B)** highlighting particular features of the nervous system. In **(A)** self-closed arrows distinguish an operation closure denoting a certain level of independence of its internal operations. Bidirectional arrows for each self-closed arrows denote the local interactions at each level (cell, system, organism), highlighting the relevance of local interactions. The cohesive and codependent interactions that lead to the emergence of a system as a recognizable unit are represented by a bigger circle denoting a self-closed arrow containing those local interactions (i.e., System, Organism). These systems may also interact with each other at a higher level (in terms of space and time) of interactions. Finally, the codependence of these systems will finally produce what we recognize as an organism following the same rationale, while then interactions are established with what we traditionally recognize as the environment (depicted by the upper curve). In the case of **(B)**, we highlighted the sensorimotor loop (internal and external), as well as the structure of sensorial structures, middle structures (traditionally referred to as processing related areas), and finally effectors. In this figure, we highlighted only muscles, but effectors cover a major range of structures, such as glands.

omits the evolutionary process that led to those functions, while also neglecting the survival of the cells that live in the organism. It is critical to note that many of our cells die each day and that each of these cells has different survival requirements and may not act in alignment with the survival of the whole organism. This is evident in pathologies such as cancer (Chaffer and Weinberg, 2011) and autoimmune diseases (Park and Kupper, 2015). We tend to refer to these conditions as errors or problems of specific systems and functions, overlooking that, since cells live in us, but not for us, there is a possibility that these phenomena may occur. As far as the global system (i.e., organism) meets its requirements, codependence relations will keep the system alive, regardless of other local interactions with no adaptive nor maladaptive values that may emerge.

Our alternative approach would be to consider that each cell meets its own requirements to survive. In this sense, it is essential to assume that the cell, as an autopoietic unit, can respond and exert control over its niche, but only within its local environment. Thus, specific environmental conditions that occur in localized regions of our body will set in motion different cellular mechanisms. Since cells can only directly influence that local environment, they can only meet their requirements. Naturally, these local interactions may have distal impacts (as Ernst Mayr conception); most of the time, when all cells meet their requirements, they indirectly end up meeting ours. As such, behavior can be considered an emergent property derived from the individual actions of cells that lead to their survival, and ultimately to ours. These two levels must be aligned for the whole organism to survive; however, there is a possibility of mismatch where some are neutral (without significant consequences) while others give rise to what we call pathology.

This different approach can be described as an interaction of parallel causes and requirements nested in cells and organisms, in the sense that the phenomena present in individual cells mirror a distal effect on the whole organism and *vice versa*. Therefore, we may explain behavior from the viewpoint of the entire organism or the interactions of its cells. However, a more comprehensive approach would be to track cellular interactions up to the mirrored effect on the organisms without neglecting that the proximal causes affecting each layer or level are aligned for survival. As such, the same phenomenon can present a

different impact on the organism and the cell populations within. For instance, covering the head with the limbs to block a hit to the head is adaptive for organisms, yet limb cells will die as a result, and the behavior would not be adaptive for them.

At this point, we may start asking which is the most relevant aspect of cells' survival. Naturally, energy management is critical for survival in any cell, as they must balance expenditure with income and maintain an adequate reserve to cope with environmental restrictions. If we consider a cell that lives within an organism and that has an evident impact over its behavior, such as a neuron, this premise stands. For neurons to survive, they must properly manage their energy budget. Problems of the organism, such as avoiding injuries, coupling with stressful work, dealing with the death of a close one, and so on, are not part of the proximal phenomena stressing a single neuron. Of course, those phenomena have stimuli transduction into local neural requirements: energy demand imposition. Therefore, neurons will deploy mechanisms to couple with their local requirements and, hopefully, they will solve the organism's problems as well. As such, when describing how behavior emerges, we should always map the differences between the organism and cellular domain of interactions. For instance, Vergara et al. (2019) described perceptual stimuli as mapped into physiology with different impacts at each level. An organism may just be looking at something. At the same time, transduction sets electromagnetic waves of the visual spectrum into action potentials, which in turn produce a cascade effect all over the nervous system, impacting the energetic demands of neurons and glia (Vergara et al., 2019). Depending on how demanding this stimulus is energetically, neurons may regulate their synaptic weights (Barral and Reyes, 2016), producing a new functional network. This new functional network will, in turn, activate muscles leading to visible behavior that may change the stimulus (e.g., closing the eyes).

This rationale is what we depict in **Figure 1B**, where we remark the particular conditions of the nervous system. All sensory inputs, driven from sensory organs, internal or external, are activated by stimulation that impose energetic demands on the nervous system. The system can affect that energy imposition by effector activity, such as muscle activation, among others. As such, closing the eyes will reduce the amount of spent energy driven by visual perception. It is also relevant to notice that for a single neuron, or even for a central nervous system neural network, it makes almost no difference if the signal arrives from interoceptive receptors or perception organs. The stimulation received is, in physical and chemical terms, the same. However, as previously implied in point 11, the feedback loops established by the nervous system through perceptual and interoceptive structures are hierarchically organized in such a way that their respective dynamics get coordinated. Also, we must not forget that the organisms not only interact with the environment through the nervous system, and that the nervous system is also coupled with other physiological systems, obeying the same rationale of local interactions.

In this framework, the energy balance mechanisms of the cells have a consequential impact on physiology resulting in the emergence of behavior. At the same time, since the cellular and whole-organism levels are analogous to nested layers or levels, the behavior itself will impact not only the experience of the whole organism but also the cells that compose it. It may be the case that only some of the cells are affected, which remarks the need of recognizing that the same phenomena may impact differently the whole organism and regions (cells) within. Also relevant is the fact that neurons cannot directly experience the stimuli that trigger organism behaviors. Once sensorial transduction is made, only proximal phenomena such as action potential, lactate transporter activation, synaptic modulations, and so on, are observable. In other words, cells such as neurons are never solving a mathematical problem, or recognizing a face, but are only solving energy needs required for their survival.

REINTERPRETING NEURAL PROCESSING

The notion that behavior is not inside the machine is notably exemplified in the experiments in "synthetic psychology" of Braitenberg (1986). He presented how simple mechanisms may lead to complex behaviors and the illusion of complex cognitive processing. The complexity may be loaned from the environment, while internal mechanisms can stay simple. We usually think of neural mechanisms as complex and difficult to assess, based on the complexity of behavior. Let us assume for a moment that it might be the case that neural mechanisms are relatively simple and that most of the complexity we see in our behavior is loaned from our environment. Is there an experiment like Braitenberg's, in which we can test real neurons?

Novellino et al. (2007) and Tessadori et al. (2013) presented an experiment resembling Braitenberg's vehicles using neuron cultures (actual neurons, not artificial neural networks). In this setup, a cart decodes distance to objects using a firing rate paradigm, and then the same paradigm is used to code back the wheels' speed independently. If the cart crashes, a stimulation burst of 20 Hz for 2 s is delivered (Tessadori et al., 2013). Under this protocol, the neuron culture learns to avoid obstacles. Thus, as external observers, we may be tempted to say that the cart does not like to crash, and it, therefore, learns to avoid obstacles. Even more, we are tempted to say that the goal of such behavior is to avoid crashes. However, that stimulation pattern is known to trigger plasticity (Madhavan et al., 2007; Chiappalone et al., 2008; le Feber et al., 2010). We may also argue that each time the cart crashes, it induces plasticity, changing the functional network. Considering how the experiment is set up, the changes will keep occurring unless crashes are avoided. Once no more crashes occur, no more changes in the network are expected. In other words, a functional neural network will keep changing until an "obstacle avoidance" structure emerges, and we will be tempted to say that the neural culture learned to avoid obstacles.

Critically, the functional network does not appear by means of an impact-avoidance goal, but as an effect derived from the energy demands posed by the stimulation that drives plasticity. Our proximate cause was energy demands, while the distal effect was avoiding obstacles. Importantly, this effect is structurally determined by how the wiring and stimulation conditions were set to the vehicle controlled by the neuron culture, meaning that a wider set of "learnings" can emerge if the structure

changes. Under this framework, it is rather useless to think that, at the neuron level, a particular neuron or set of neurons are "processing obstacle avoidance", or that there is an obstacle avoidance network in the neuronal culture. At the level of the organism, we can be tempted to use this approach, and it might be even helpful in some contexts. Nonetheless, to explain how the vehicle learns, we must consider that individual neurons deal with significant energy demands that trigger plasticity as a compensation mechanism (Vergara et al., 2019), which produces the avoidance of obstacles as emergent behavior.

It is possible to establish that, in proximal terms, neurons must efficiently solve their energy management. As depicted in **Figure 2**, we expect that a neural network in equilibrium will lose its energy balance driven by external stimuli. The energy imbalance will propagate through the network according to its structural constraints. Since most neural connections with different regions are bidirectional, the system will generate a global answer (as observers, we may declare it a coordinated answer). Eventually, this will get to the effectors (full propagation is achieved). At that moment, the organism will be able to take action as a whole system to impact the input stimulus that has disturbed the energy balance. It is critical to notice that, in the meantime, local mechanisms of single neurons are triggered to couple with this increment in energy demand as well. Within this close-loop structure, the actions taken by the entire organism, as well as those taken by individual cells, will allow a new energy equilibrium to be achieved, which will be a novel functional neural structure associated with a novel behavior.

From the previous argument highlighting energy as a key regulatory element, makes sense considering neurons' proximal context as the trigger of neuron regulation. Neurons are extremely sensitive to oxygen deprivation (Ames, 2000) and the central nervous system possesses small glycogen reserves (Brown and Ransom, 2007). Neurons answer to energy demands (neural activity) by outsourcing their energy needs to the glia (Weber and Barros, 2015), which will trigger the neurovascular coupling associated with neural activity (Sokoloff, 2008; Schulz et al., 2012; Robinson and Jackson, 2016), followed by increased glucose uptake and glycolytic rate of astrocytes (Magistretti and Allaman, 2018). In addition, neuronal mitochondria increase ATP synthesis in response to an increment in synaptic stimuli (Jekabsons and Nicholls, 2004; Connolly et al., 2014; Rangaraju et al., 2014; Toloe et al., 2014; Lange et al., 2015). These are just the early responses in the range of hours, as the synaptic scaling ends balancing to a homeostatic level of neurons' activity (Barral and Reyes, 2016), reducing the energy cost of the activity increment. An increment in stimulation is expected to produce long-term network modularization (Novellino et al., 2007). Interestingly enough, when significant downscaling occurs, a few synaptic weights (dendritic spines) will increase (El-Boustani et al., 2018; Jungenitz et al., 2018). As such, while we have only described the proximal actions of neurons, they have a vast impact on the neural networks and therefore behavior. It is plausible to observe neural processing as an emergent property rooted in proximal cells requirements.

Up to this point, we have been able to rephrase neural processing without purpose or so-called "goal-oriented behaviors". Our explanation has also been faithful to a structural determinism, meaning that behavior in the vehicle (i.e., neuronal culture) emerges as a result of neurons doing the actions determined by their properties and structure. So far, introducing volition or desires in this context would be to acknowledge openly that a culture of neurons shares the same properties we usually attribute to a whole organism. However, does this reinterpretation lead to new implications?

The most obvious is the reinterpretation of key phenomena into local interactions. For instance, neural activity, usually seen as neural processing, would be interpreted as energy expenditure, as an environmental pressure for a neuron, which forces it to activate mechanisms to balance its energy budget. Otherwise, it dies. Plasticity, classically viewed as a learning mechanism (please note how a molecular mechanism has a whole system property; learning), would be reinterpreted as a coupling mechanism of neurons to deal with incoming energetic demands from presynaptic neurons. As stated in Vergara et al. (2019), the synaptic gain will change to match a homeostatic energy level. This immediately sets some empirical implications. For instance, synaptic scaling should answer to stimulation, but also to energy availability. Therefore, changes in glucose availability in a neuronal culture should change the dynamics of classic synaptic scaling protocols. Specifically, synaptic scaling should be higher in the case of less glucose availability (for more empirical predictions, see Vergara et al., 2019).

Another consequence of viewing neuronal processing as an emergent property of individual neurons displaying mechanisms that allow them to stay alive under different energy pressures is that not everything neurons do is helpful to the organism. In other words, since neurons are only solving their local requirements, their actions may lead to the emergence of useless behaviors. This means that part of the neural network activity, which can respond to the continuous activity of multiple stimuli, will lead to the appearance of behaviors with no apparent usefulness and that may even be maladaptive. This is necessarily the consequence of a codependent system governed by local actions. Each component solves its requirements as part of its condition of existence, but once they are solved, other harmless actions may occur as a kind of debris that results from the operational closure. It is crucial to notice that as living beings we do not need a perfect functional coupling with the environment; it must be just good enough to survive. If we consider further the hierarchic structure, even relevant actions for survival at a single cell level may have useless or undesired impacts at the whole-organism level. As long as survival is not immediately compromised, e.g., as far as physiologically critical homeostatic systems are not driven away from equilibrium, these mismatches may freely occur. This consequence frees us from the need to include a function in every behavior we have. Many of them can be helpful for our survival and others may not, but above all, given the degree of freedom allowed by the hierarchical organization of the neural-body-environment homeostatic mechanisms, we can have neutral behaviors from an adaptative viewpoint.

This last point is critical since the degree of behavioral flexibility increases the probability of producing neutral

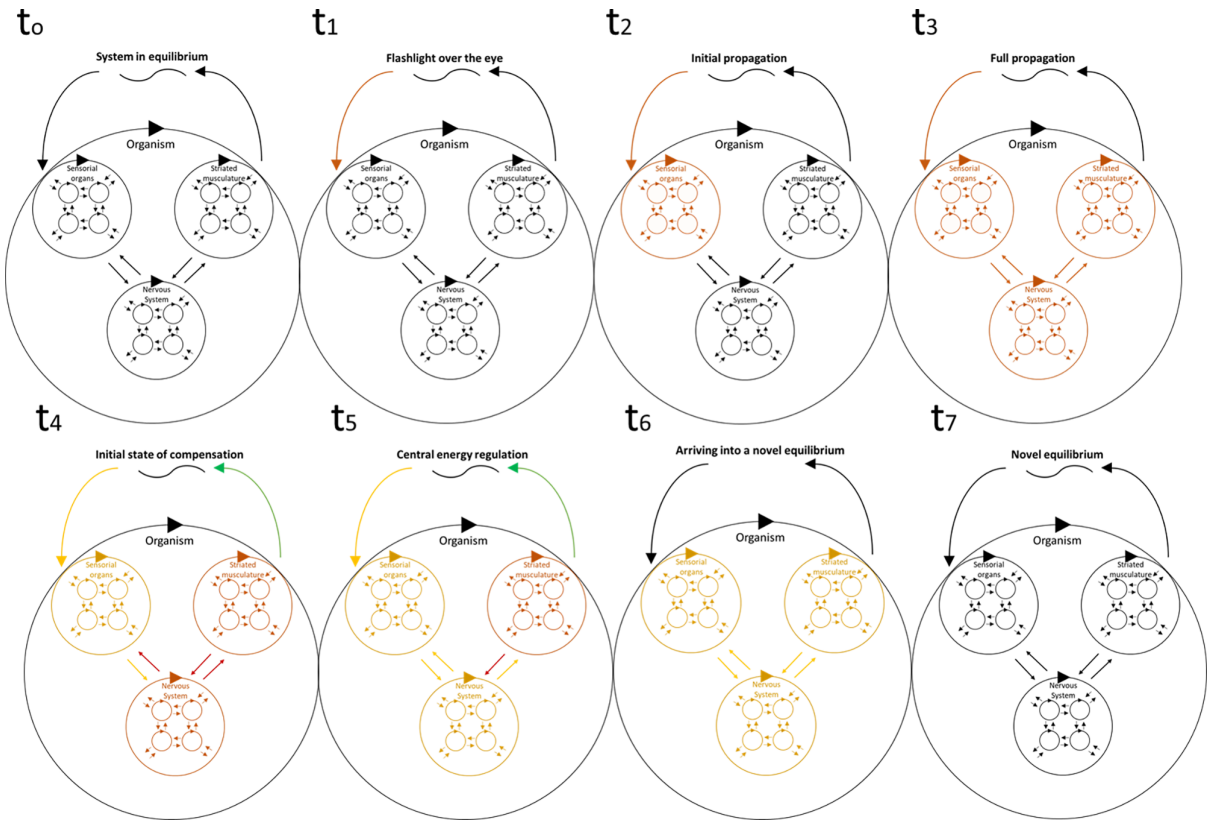


FIGURE 2 | A simplified version of **Figure 1B** depicting temporal dynamics driven by external stimuli that push the system out of its energetic equilibrium. This figure does not intend to be an exhaustive description of the temporal dynamics of this process, rather to depict some critical features of such dynamics. In particular, that the energy demand imposition of the initial stimulus would only be detected by the sensory regions, and then propagate into the remaining areas. We also emphasize that these regions are bidirectionally related, so that, before the behavior occurs, it is expected that some energy compensations of single neurons and neural networks will take place. When the system is in full propagation of the energy demand imposed by the environment, a behavior will take place impacting the perceived environment. This will have an effect on the sensory areas. However, the impact is not immediate, and it will return to a propagation process in which the mechanisms of all levels (neuron, neural networks, organism) will find a novel equilibrium.

behaviors and deleterious ones. Thus, it is not surprising that animals with high behavioral flexibility are associated with greater effort and parenting times during ontogeny (Isler and van Schaik, 2009, 2012; Barton and Capellini, 2011; Heldstab et al., 2019; Uomini et al., 2020). One only needs to observe how a toddler relates to its environment to discover that many of our behaviors during infancy put our survival or fitness at peril. Parental care or parenting allows us to buffer this flexibility, allowing us to stay alive. Conversely, flexibility also allows us to increase fitness by adapting to the environment during ontogeny, unlike less flexible animals requiring phylogenetic mechanisms of change to adapt.

BUILDING UP TO COMPLEX BEHAVIORS

Behavioral flexibility by means of EHP is a powerful concept, as it explains fast changes in behavior during ontogeny, but it also allows the test-retest rationale to operate. As far as the test-retest rationale follows the restrictions imposed by single-cell energy management, learning can emerge. We expect that this flexibility is what ultimately gives rise to the most complex cognitive phenomena, such as understanding. Specifically, what we refer to as useless behavior can be interpreted out of the teleological paradigm as behavioral flexibility. Those apparently useless behaviors may find their usefulness when an environmental pressure is relieved by this behavior, or they may never find their usefulness from the observer's position. From a naturalistic approach, this is just flexibility to couple with the environment following point 8, describing the nervous system as a homeostatic system that will maintain certain stability and equilibrium and restore it to a certain extent.

In this view, complex cognitive phenomena emerge from this hierarchical flexibility of the system. These more sophisticated cognitive phenomena are vastly discussed and modeled using the Free Energy Principle (Friston, 2010). How does EHP stand in contrast to FEP? The FEP is an organism-based approach that considers volition as a critical element, especially when regarding aspects such as understanding (Yufik and Friston, 2016), as it distinguishes lower forms of learning, allowing the introduction of cognitive models. Therefore, as an initial difference, we noticed that FEP rather omits neuron requirements, assuming them as chronically met. Secondly, it

assumes the presence of goal-oriented behaviors, volition, and purpose, which is to be expected if starting from a whole organism viewpoint.

Although there are obvious differences between these two perspectives, especially since the FEP contains teleological elements and considers the nervous system as an epistemic agent (points 1 and 3 of "General Systemic Conditions" section), Yufik's proposal (Yufik, 2013, 2019; Yufik and Friston, 2016) is very similar to the EHP at the neural network level. He developed the idea of how neural assemblies (or packets) would appear, producing functional networks which allow understanding to emerge. The core idea envisions the mind as a cartographer mapping the environment, similar to classic cognitive perspectives (Bateson, 2015), where modularization of functional neural activity will allow differences to be made. In order to establish that two objects are different, a difference in the functional network should emerge (different packets or sets of packets) to allow the recognition of such distinction. Following our rationale, the critical question is what local mechanism is driving the emergence of those distinctions. During the works of this thermodynamic conception of cognition, there is an acknowledgment of the relevance of energy in modulating the packets' emergence in this proposal (Yufik, 2013, 2019; Yufik and Friston, 2016). For instance, cortical tone (temperature of this thermodynamic formalization), which can be rephrased as energy demands using EHP, is critical in how the system will react towards the equilibrium by FEP conception (Yufik, 2013).

This is a critical aspect, as in this FEP-driven proposal energy conditions modify the neural functional structure to produce a novel equilibrium. This conceptually very similar to the EHP, as depicted in **Figure 2**, achieves a novel equilibrium by a new energetic demand (i.e., cortical tone). Even more interesting is that in Yufik's work (Yufik, 2013, 2019; Yufik and Friston, 2016) modularization is expected from a learning process, the same process reported by Tessadori et al. (2013) and Novellino et al. (2007), which we have explained from an EHP viewpoint above. This role of energy management is even more explicit in the following communications (Yufik, 2019), mainly focused on the demand or energy expenditure and availability. Therefore, both approaches find common ground in the middle, acknowledging that local neuron requirements (i.e., energy management) are critical for modularization to occur, leading to cognitive distinctions that will ultimately produce understanding.

It is relevant to notice that EHP and FEP are two sides of the same coin. Following the parallel conception of organism vs. cell community approach, all conceptions derived from FEP could be mapped in EHP terms and *vice versa*. Naturally, as we get closer to cellular processes, FEP is less precise on its implications, and when getting closer to high cognitive functions, EHP is rather vague. However, reasonable efforts can be made to understand what is happening at cellular and physiological levels when we describe the cognitive mechanism. For instance, one challenging explanation to be made from the EHP side is anticipatory behavior. How can neurons caring about their local needs solve upcoming organism events?

One key aspect of anticipatory behavior is that it must be learned first. In other words, it is not anticipating anything, it is rather re-evoking structural history. This means that most predictions we make are based upon past experiences. Therefore, we avoid pain, as we have previously experienced pain. Similar to what we described in Tessadori's vehicle case (Tessadori et al., 2013), energy demands derived from the painful stimulation lead to restructuration, allowing pain avoidance to occur (rephrased as reducing surprise by FEP means). If we focused not on the result but on the learning phase, we would notice that consistent unrelated stimulus (e.g., a light turning on, an acoustic event, or a similar signal event) is followed by pain.

Light, sound, and pain produce energy demands through perception. Nonetheless, the pain has a durable effect, which means a long-lasting energy demand situation. Also, its intensity is directly related to the amount of damage (Dubin and Patapoutian, 2010). Therefore, that is the critical stimulation to be avoided by means of local neuron requirements.

When we focus on neural activity during situations of these characteristics, we observe that both neural activities, the one derived from the upcoming pain signal and the one directly derived from pain, begin to fire closer in time through learning (Urien et al., 2018). The overall activity appears to be the same, but the temporal aspect change. Basically, now the signal triggers both the signal-related activity and the originally pain-driven avoidance behavior. The critical aspect here is that the signal that anticipates pain does not mean pain itself, but in neural activity, the signal packet (assembly) will fire just before the avoidance behavior packet. Following the logic of *fire together, wire together*, the avoidance packet will ultimately be activated without the pain but with the signal packet, meaning the fusion of these two packets. Please note that this explanation does not involve mental manipulations yet as the ones suggested by FEP, and we can still be faithful to our premises.

From EHP, the fusion of these neural activities into one module that would lead to the so-called anticipatory behavior, is driven by the same rationale observed in **Figure 2**. Basically, the initial trial will deploy many behaviors that will not be useful to keep the equilibrium, while at the same time the propagation of the energy demand imposed by pain will, in consequence, functionally restructure the network with each iteration. Following the same proposed mechanism for the vehicle controlled by a neuronal culture, at some point behavior will satisfy the condition of approaching neurons to a novel equilibrium. During this central energy regulation, with each iteration the best "pain-avoidance" structure will be selected until the predictive behavior is settled. These changes may even follow a random structure change, and they would still work. However, neural mechanisms such as synaptic reinforcement by fire together, wire together (Abbott and Nelson, 2000), play a critical role for this to happen efficiently. Considering that the EHP reinterprets these plastic mechanisms as coping energy mechanisms of neurons, we are able to explain these phenomena without yet needing to call for complex mental scenarios. Naturally, this explanation does not cover more sophisticated behaviors like planning, which under a classic view require volitional manipulation of information. However, it sheds light

on how, starting from cellular communities, "goal-directed" behaviors can be explained leaving the goal as the consequence, not the cause. Neurons don't even realize that the animal was submitted to pain; they just react according to their local requirements. It is we who, as observers, are tempted to say that the animal learns to anticipate the aversive stimulus. Even more relevant is the fact that as we show anticipatory behavior, we may be blind to the actual causes that led to this apparent anticipatory behavior by neglecting history, which under the EHP view is no more than an expression of an organism coupled with its environment where its particular history defines the behaviors that will be deployed when observing the signal related to pain.

Under this context, we have given an explanation of how an organism can act in the prediction of hazard, without actually predicting it. Local neural properties allow these phenomena to occur without incorporating purpose, mental model, or further mental scenarios. Notably, FEP and EHP, despite their differences in starting points (and, therefore, conceptual frameworks), share similar predictions on how neural networks would operate. Distinctions are made on what produces those changes. Another relevant difference of our approach is that neurons can fulfill their requirements without solving the problem of the whole organism but never endangering the life of the organism (at least not immediately). Therefore, the behavioral flexibility given by the impact left by neurons when solving their needs could have a negative, neutral, or positive impact, which means that the neurons may find local energy homeostasis attractors that satisfy their requirements but not necessarily the organism's requirements. However, if so, why does it seem that they are almost always positive (hence the teleological need to indicate their function)?

AN EVOLUTIONARY PERSPECTIVE ON THE COUPLING OF DIFFERENT LEVELS OF OPERATIONAL CLOSURES

We see what remains, not what has been. During the evolutionary history of living beings, most species have disappeared, have become extinct (Newman, 1997). In fact, the species that are alive today represent less than 1% of the historical total (Newman, 1997; Jablonski, 2004). This makes it risky to use evidence only from modern animals to explain the relationship between the cellular and whole-organism levels of organization. On the other hand, virtually all present-day animal body plans date at least back to the Cambrian explosion (CE), an event that occurred more than 500 million years ago (Maloof et al., 2010). While it is still a matter of debate, it is possible to propose that near that time window, a level of animal diversification and radiation occurred that had not been seen before and has not been seen since (Keijzer, 2015; Trestman, 2013).

Interestingly, this period has also been ascribed as when metazoans with complex active bodies appeared (Trestman, 2013). These organisms are defined by having: (i) articulated appendages; (ii) many degrees of freedom of controlled movement; (iii) true senses (with specialized organs such as eyes); (iv) sense-guided motility; and (v) anatomical capacity for object manipulation (Trestman, 2013). The appearance of metazoans probably occurred at least 200 million years before the CE (Erwin, 2015; Dohrmann and Wörheide, 2017), and the nervous system probably appeared during the Ediacaran period (635 million years ago). In simpler metazoans with low-complexity nervous systems, synchrony between the neuronal and organism levels was probably much easier to achieve than in animals with complex active bodies. Movement is not yet a problem for those animals. Thus, it is feasible that, during the initial evolution of the nervous system, a limiting element was the alignment between the neuronal level and that of the whole organism. Once this occurred, the space for possible radiation and diversification opened up.

In ontogenetic terms, the reality is similar. In animals, the highest mortality rates are usually seen early in life (Caughley, 1966), when their individual-environment relationships are still being established and they tend to have much more behavioral flexibility. Even in our species, this reality is not far off, for it has not been long since most of our offspring died during the first 3 years of life (Volk and Atkinson, 2013). The problem lies in that we often only consider its present condition when observing an organism such as ourselves and its direct relationship with the environment, ignoring its phylogenetic and ontogenetic history. Under this perspective, most cellular phenomena are aligned with their whole-organism functions. This may lead to the interpretation that the proportion of misaligned events between these levels of the organization is negligible or almost nonexistent. Thus, we only see what has worked for survival, while counterexamples of instances where cellular phenomena are misaligned with organisms vanish. In other words, under this view, we are incurring a survival bias, where we focus only on the instances where cellular and whole-organism levels overcame a selection process and overlook those that did not. This can lead us to false conclusions, such as overrepresenting aligned states or assuming cellular levels have functions for our survival.

This also translates into a tradeoff between flexibility and survivability. Higher degrees of freedom and higher levels of flexibility allow the emergence of novel adaptations, which increase the organism's fitness. This context can also explain why larger nervous systems (brains with more neurons) are associated with greater behavioral richness. A larger number of neurons leads to a greater diversity of local responses/solutions and greater behavioral flexibility. However, on the other hand, there may be a maximum of possible degrees of freedom before the number of misalignments between cellular and whole-organism levels can remain functional.

Another point to consider is that not necessarily every lack of synchrony is maladaptive. There is the possibility that some of the neuronal activity that is not fully aligned with the organism is "neutral." Thus, analogous to models of neutral evolution, it is feasible that a non-trivial proportion of what neurons do to solve their local energy requirements has no significant impact on the organism's survival. It is possible to postulate that the less fundamental to survival a behavior is, the more neutral activity there is. That is, the less essential behaviors probably allow for

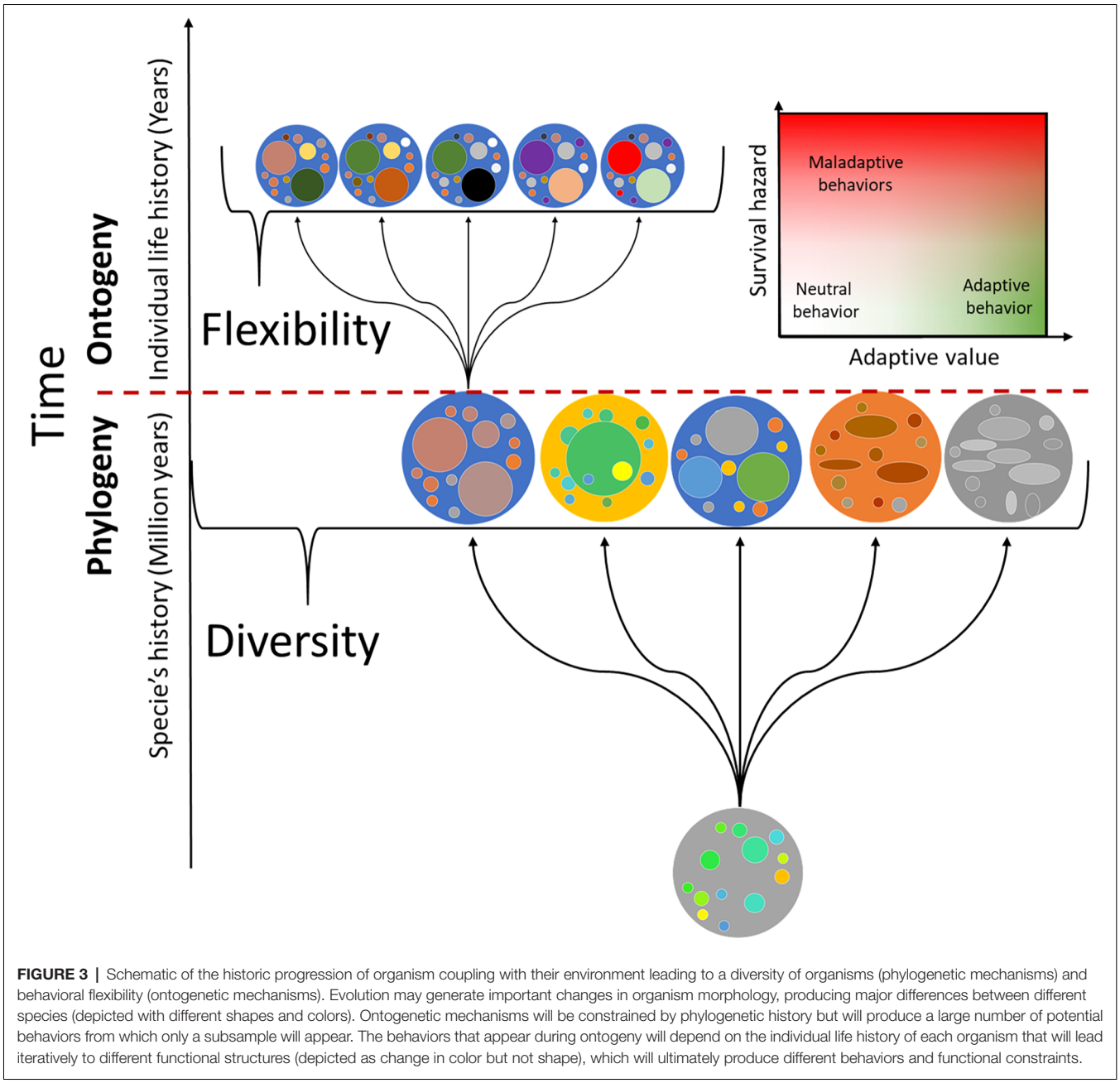


FIGURE 3 | Schematic of the historic progression of organism coupling with their environment leading to a diversity of organisms (phylogenetic mechanisms) and behavioral flexibility (ontogenetic mechanisms). Evolution may generate important changes in organism morphology, producing major differences between different species (depicted with different shapes and colors). Ontogenetic mechanisms will be constrained by phylogenetic history but will produce a large number of potential behaviors from which only a subsample will appear. The behaviors that appear during ontogeny will depend on the individual life history of each organism that will lead iteratively to different functional structures (depicted as change in color but not shape), which will ultimately produce different behaviors and functional constraints.

less alignment between levels. This, in turn, would increase the presence of behavioral richness or "polymorphisms" in those behaviors. Specifically, when both the adaptive value and the survival hazard are low, neutral behavior emerges (**Figure 3**).

Finally, it is critical to realize that, under this notion, behaviors are not goal-oriented *per se*. Many may appear as goal-directed, as they are conditions of existence of the system (e.g., breathing). Under our scope, breathing organisms stay alive, therefore exist. However, breathing was never designed or deliberately addressed to meet the oxygen requirements of the organism. When we remove the goal rationale of structures and behavior, the evolutive process in which behavior emerges loses its need for teleological explanation. As such, the brain or areas within were not designed to solve specific problems. Instead, in meeting their own requirements, cells satisfy the organism's requirements too; if not, survival is compromised. When most cells living in the cellular community meet their requirements, the organisms will do so. It is simply the condition of existence of such a community. Behavioral diversity and flexibility emerge within these messy interactions of individual cells acting locally and producing distal effects that may not even affect them directly.

FINAL REMARKS

When we observe a single cell acting in an anticipatory fashion (e.g., Shirakawa, 2006), we avoid attributing it to a sense of

volition, or any epistemic or informational operation. We focus on its local mechanisms which result in such anticipatory behavior. Avoiding it is reasonable, as including it obscures the mechanisms, and we also recognize the cell as a physical system determined by the mechanisms governing it. For some reason, when coming to human beings, we fail to recognize them in such a way. This is so dramatic, that besides EHP, we have no knowledge of another integrative explicative proposal of behavior using a strict naturalistic approach.

FEP is probably the most sophisticated and flexible proposal explaining human behavior as an integrative framework. However, it uses a strong epistemic rationale to explain behavior. This leads to assigning volition to all living beings (or even dissipative systems) or stating that the concept is only applicable for certain systems such as human beings. Despite the differences, it is notable that the phenomena described at the neural network level are quite similar in both proposals, meaning that both recognize more or less the same events as relevant to explain behavior. The causes of those events are different depending on which proposal framework is used.

We understand that intending to explain behavior and most sophisticated forms of it, such as understanding, is a major challenge for EHP. However, we consider that it is a required academic exercise in our current framework of neuroscience. As we have stated above, goals can easily emerge as observer assignation once the system is coupled with its environment, but from an evolutionary perspective, adaptations do not appear to solve a problem; they just appear, and they are preserved due to advantageous (or at least non-deleterious) impacts. In other words, focusing on the goal may obscure the actual mechanisms that produce the phenomena we look forward to understanding.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SV-J, MV, and RV developed the initial general argument. All authors contributed to all the drafts; nonetheless, each author did special contributions to different sections. MV formalized the general systemic conditions. RV contributed in the neural processing and building up to behavior sections. SV-J developed the evolutionary perspective. RV and PM edited the final version. All authors contributed to the article and approved the submitted version.

FUNDING

MV's contribution to this article was supported by the Agencia Nacional de Investigación y Desarrollo, ANID, from a grant FONDECYT REGULAR 1191477. PM's contribution to this article was supported by Proyect ICN09_015.

REFERENCES

Abbott, L. F., and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3, 1178–1183. doi: 10.1038/81453

Ames, A., 3rd (2000). CNS energy metabolism as related to function. *Brain Res. Rev.* 34, 42–68. doi: 10.1016/s0165-0173(00)00038-2

Ashby, W. R. (1947). The nervous system as physical machine; with special reference to the origin of adaptive behaviour. *Mind* 56, 44–59. doi: 10.1093/mind/lvi.221.44

Ashby, W. R. (1960). *Design for a Brain*, 2nd edition. London: Chapman & Hall.

Barral, J., and Reyes, A. D. (2016). Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics. *Nat. Neurosci.* 19, 1690–1696. doi: 10.1038/nn.4415

Barton, R. A., and Capellini, I. (2011). Maternal investment, life histories and the costs of brain growth in mammals. *Proc. Natl. Acad. Sci. U S A* 108, 6169–6174. doi: 10.1073/pnas.1019140108

Bateson, G. (2015). Form, substance and difference. *ETC: A Review of General Semantics* 72, 90–104. Available online at: http://www.jstor.org/stable/24761998.

Bechtel, W., and Richardson, R. C. (1998). "Vitalism," in *Routledge Encyclopedia of Philosophy*, ed E. Craig (London: Routledge), 639–643.

Bonabeau, E., Theraulaz, G., Deneubourg, J. L., Aron, S., and Camazine, S. (1997). Self-organization in social insects. *Trends Ecol. Evol.* 12, 188–193. doi: 10.1016/s0169-5347(97)01048-3

Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology.* Cambridge, MA: MIT Press.

Brown, A. M., and Ransom, B. R. (2007). Astrocyte glycogen and brain energy metabolism. *Glia* 55, 1263–1271. doi: 10.1002/glia.20557

Caughley, G. (1966). Mortality patterns in mammals. *Ecology* 47, 906–918. doi: 10.2307/1935638

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Chiappalone, M., Massobrio, P., and Martinoia, S. (2008). Network plasticity in cortical assemblies. *Eur. J. Neurosci.* 28, 221–237. doi: 10.1111/j.1460-9568.2008.06259.x

Connolly, N. M. C., Dussmann, H., Anilkumar, U., Huber, H. J., and Prehn, J. H. M. (2014). Single-cell imaging of bioenergetic responses to neuronal excitotoxicity and oxygen and glucose deprivation. *J. Neurosci.* 34, 10192–10205. doi: 10.1523/JNEUROSCI.3127-13.2014

Dada, J. O., and Mendes, P. (2011). Multi-scale modelling and simulation in systems biology. *Integr. Biol.* 3, 86–96. doi: 10.1039/c0ib00075b

DiFrisco, J. (2017). Time scales and levels of organization. *Erkenntnis* 82, 795–818. doi: 10.1007/s10670-016-9844-4

Dohrmann, M., and Wörheide, G. (2017). Dating early animal evolution using phylogenomic data. *Sci. Rep.* 7:3599. doi: 10.1038/s41598-017-03791-w

Dubin, A. E., and Patapoutian, A. (2010). Nociceptors: the sensors of the pain pathway. *J. Clin. Invest.* 120, 3760–3772. doi: 10.1172/JCI42843

El-Boustani, S., Ip, J. P. K., Breton-Provencher, V., Knott, G. W., Okuno, H., Bito, H., et al. (2018). Locally coordinated synaptic plasticity of visual cortex neurons *in vivo*. *Science* 360, 1349–1354. doi: 10.1126/science.aao0862

Engel, G. L. (1980). The clinical application of the biopsychosocial model. *Am. J. Psychiatry* 137, 535–544. doi: 10.1176/ajp.137.5.535

Erwin, D. H. (2015). Early metazoan life: divergence, environment and ecology. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20150036. doi: 10.1098/rstb.2015.0036

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Gillooly, J. F., Hou, C., and Kaspari, M. (2010). Eusocial insects as superorganisms: insights from metabolic theory. *Commun. Integr. Biol.* 3, 360–362. doi: 10.4161/cib.3.4.11887

Heldstab, S. A., Isler, K., Burkart, J. M., and van Schaik, C. P. (2019). Allomaternal care, brains and fertility in mammals: who cares matters. *Behav. Ecol. Sociobiol.* 73:71. doi: 10.1007/s00265-019-2684-x

Herbert-Read, J. E. (2016). Understanding how animal groups achieve coordinated movement. *J. Exp. Biol.* 219, 2971–2983. doi: 10.1242/jeb.129411

Herbert-Read, J. E., Perna, A., Mann, R. P., Schaerf, T. M., Sumpter, D. J., and Ward, A. J. (2011). Inferring the rules of interaction of shoaling fish. *Proc. Natl. Acad. Sci. U S A* 108, 18726–18731. doi: 10.1073/pnas.1109355108

Isler, K., and van Schaik, C. P. (2009). Why are there so few smart mammals (but so many smart birds)? *Biol. Lett.* 5, 125–129. doi: 10.1098/rsbl.2008.0469

Isler, K., and van Schaik, C. P. (2012). Allomaternal care, life history and brain size evolution in mammals. *J. Hum. Evol.* 63, 52–63. doi: 10.1016/j.jhevol.2012.03.009

Jablonski, D. (2004). Extinction: past and present. *Nature* 427:589. doi: 10.1038/427589a

Jekabsons, M. B., and Nicholls, D. G. (2004). in situ respiration and bioenergetic status of mitochondria in primary cerebellar granule neuronal cultures exposed continuously to glutamate. *J. Biol. Chem.* 279, 32989–33000. doi: 10.1074/jbc.M401540200

Jungenitz, T., Beining, M., Radic, T., Deller, T., Cuntz, H., Jedlicka, P., et al. (2018). Structural homo- and heterosynaptic plasticity in mature and adult newborn rat hippocampal granule cells. *Proc. Natl. Acad. Sci. U S A* 115, E4670–E4679. doi: 10.1073/pnas.1801889115

Keijzer, F. (2015). Moving and sensing without input and output: early nervous systems and the origins of the animal sensorimotor organization. *Biol. Philos.* 30, 311–331. doi: 10.1007/s10539-015-9483-1

Lange, S. C., Winkler, U., Andresen, L., Byhrø, M., Waagepetersen, H. S., Hirrlinger, J., et al. (2015). Dynamic changes in cytosolic ATP levels in cultured glutamatergic neurons during NMDA-induced synaptic activity supported by glucose or lactate. *Neurochem. Res.* 40, 2517–2526. doi: 10.1007/s11064-015-1651-9

le Feber, J., Stegenga, J., and Rutten, W. L. C. (2010). The effect of slow electrical stimuli to achieve learning in cultured networks of rat cortical neurons. *PLoS One* 5:e8871. doi: 10.1371/journal.pone.0008871

Madhavan, R., Chao, Z. C., and Potter, S. M. (2007). Plasticity of recurring spatiotemporal activity patterns in cortical networks. *Phys. Biol.* 4, 181–193. doi: 10.1088/1478-3975/4/3/005

Magistretti, P. J., and Allaman, I. (2018). Lactate in the brain: from metabolic end-product to signalling molecule. *Nat. Rev. Neurosci.* 19, 235–249. doi: 10.1038/nrn.2018.19

Maloof, A. C., Porter, S. M., Moore, J. L., Dudas, F. O., Bowring, S. A., Higgins, J. A., et al. (2010). The earliest Cambrian record of animals and ocean geochemical change. *Geol. Soc. Am. Bull.* 122, 1731–1774. doi: 10.1130/B30346.1

Maturana, H. (1978). "Cognition," in *Wahrnehmung und Kommunikation*, eds P. M. Hejl, W. K. Köck and G. Roth (Frankfurt: Peter Lang), 29–49.

Maturana, H. R. (1980). "Biology of cognition," in *Autopoiesis and Cognition. Boston Studies in the Philosophy and History of Science*, eds H. Maturana and F. J. Varela (Dordrecht: Springer), 1–58. doi: 10.1007/978-94-009-8947-4_5

Maturana, H. (2002). Autopoiesis, structural coupling and cognition: a history of these and other notions in the biology of cognition. *Cybern. Hum. Knowing* 9, 5–34.

Maturana, H. R. (2008). Anticipation and self-consciousness. Are these functions of the brain? *Constructivist Found.* 4, 18–20. Available online at: http://constructivist.info/4/1/018.

Maturana, H. R., and Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boulder, CO: New Science Library/Shambhala Publications.

Mayr, E. (1961). Cause and effect in biology: kinds of causes, predictability and teleology are viewed by a practicing biologist. *Science* 134, 1501–1506. doi: 10.1126/science.134.3489.1501

McGregor, S., and Virgo, N. (2011). Life and its close relatives. *Lect. Notes Comput. Sci.* 5778, 230–237. doi: 10.1007/978-3-642-21314-4_29

Newman, M. E. J. (1997). A model of mass extinction. *J. Theor. Biol.* 189, 235–252. doi: 10.1006/jtbi.1997.0508

Novellino, A., D'Angelo, P., Cozzi, L., Chiappalone, M., Sanguineti, V., and Martinoia, S. (2007). Connecting neurons to a mobile robot: an *in vitro* bidirectional neural interface. *Comput. Intell. Neurosci.* 2007:12725. doi: 10.1155/2007/12725

Park, C. O., and Kupper, T. S. (2015). The emerging role of resident memory T cells in protective immunity and inflammatory disease. *Nat. Med.* 21, 688–697. doi: 10.1038/nm.3883

Pedersen, S. F., Kapus, A., and Hoffmann, E. K. (2011). Osmosensory mechanisms in cellular and systemic volume regulation. *J. Am. Soc. Nephrol.* 22, 1587–1597. doi: 10.1681/ASN.2010121284

Rangaraju, V., Calloway, N., and Ryan, T. A. (2014). Activity-driven local ATP synthesis is required for synaptic function. *Cell* 156, 825–835. doi: 10.1016/j.cell.2013.12.042

Robinson, M. B., and Jackson, J. G. (2016). Astroglial glutamate transporters coordinate excitatory signaling and brain energetics. *Neurochem. Int.* 98, 56–71. doi: 10.1016/j.neuint.2016.03.014

Roux, E. (2014). The concept of function in modern physiology. *J. Physiol.* 592, 2245–2249. doi: 10.1113/jphysiol.2014.272062

Schulz, K., Sydekum, E., Krueppel, R., Engelbrecht, C. J., Schlegel, F., Schröter, A., et al. (2012). Simultaneous BOLD fMRI and fiber-optic calcium recording in rat neocortex. *Nat. Methods* 9, 597–602. doi: 10.1038/nmeth.2013

Shirakawa, T. (2006). Anticipatory behavior and intracellular communication in *Physarum polycephalum*. *AIP Conference Proceedings (AIP)* 839, 541–546. doi: 10.1063/1.2216665

Sokoloff, L. (2008). The physiological and biochemical bases of functional brain imaging. *Cogn. Neurodyn.* 2, 1–5. doi: 10.1007/s11571-007-9033-x

Southern, J., Pitt-Francis, J., Whiteley, J., Stokeley, D., Kobashi, H., Nobes, R., et al. (2008). Multi-scale computational modelling in biology and physiology. *Prog. Biophys. Mol. Biol.* 96, 60–89. doi: 10.1016/j.pbiomolbio.2007.07.019

Tessadori, J., Venuta, D., Kumar, S. S., Bisio, M., Pasquale, V., and Chiappalone, M. (2013). "Embodied neuronal assemblies: a closed-loop environment for coding and decoding studies," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, (San Diego, CA, USA), 899–902. doi: 10.1109/NER.2013.6696080

Toloe, J., Mollajew, R., Kügler, S., and Mironov, S. L. (2014). Metabolic differences in hippocampal "Rett" neurons revealed by ATP imaging. *Mol. Cell. Neurosci.* 59, 47–56. doi: 10.1016/j.mcn.2013.12.008

Trestman, M. (2013). The cambrian explosion and the origins of embodied cognition. *Biol. Theory* 8, 80–92. doi: 10.1007/s13752-013-0102-6

Ulanowicz, R. E., and Hannon, B. M. (1987). Life and the production of entropy. *Proc. R. Soc. London. Ser. B. Biol. Sci.* 232, 181–192. doi: 10.1098/rspb.1987.0067

Uomini, N., Fairlie, J., Gray, R. D., and Griesser, M. (2020). Extended parenting and the evolution of cognition. *Philos. Trans. R. Soc. B Biol. Sci.* 375:20190495. doi: 10.1098/rstb.2019.0495

Urien, L., Xiao, Z., Dale, J., Bauer, E. P., Chen, Z., and Wang, J. (2018). Rate and temporal coding mechanisms in the anterior cingulate cortex for pain anticipation. *Sci. Rep.* 8:8298. doi: 10.1038/s41598-018-26518-x

Vergara, R. C., Jaramillo-Riveri, S., Luarte, A., Moënne-Loccoz, C., Fuentes, R., Couve, A., et al. (2019). The energy homeostasis principle: neuronal energy regulation drives local network dynamics generating behavior. *Front. Comput. Neurosci.* 13:49. doi: 10.3389/fncom.2019.00049

Villalobos, M. E. (2015). *Biological roots of cognition and the social origins of mind: autopoietic theory, strict naturalism and cybernetics*. University of Edinburgh, United Kingdom. Available online at: http://hdl.handle.net/1842/26004.

Villalobos, M., and Ward, D. (2015). Living systems: autonomy, autopoiesis and enaction. *Philos. Technol.* 28, 225–239. doi: 10.1007/s13347-014-0154-y

Volk, A. A., and Atkinson, J. A. (2013). Infant and child death in the human environment of evolutionary adaptation. *Evol. Hum. Behav.* 34, 182–192. doi: 10.1016/j.evolhumbehav.2012.11.007

Weber, B., and Barros, L. F. (2015). The astrocyte: powerhouse and recycling center. *Cold Spring Harb. Perspect. Biol.* 7:a020396. doi: 10.1101/cshperspect.a020396

Yufik, Y. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. (2019). The understanding capacity and information dynamics in the human brain. *Entropy (Basel)* 21:308. doi: 10.3390/e21030308

Yufik, Y., and Friston, K. (2016). Life and understanding: the origins of "understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 11 January 2022
doi: 10.3389/fnsys.2021.767461

Predictive Neuronal Adaptation as a Basis for Consciousness

Artur Luczak and Yoshimasa Kubo*

Canadian Center for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Being able to correctly predict the future and to adjust own actions accordingly can offer a great survival advantage. In fact, this could be the main reason why brains evolved. Consciousness, the most mysterious feature of brain activity, also seems to be related to predicting the future and detecting surprise: a mismatch between actual and predicted situation. Similarly at a single neuron level, predicting future activity and adapting synaptic inputs accordingly was shown to be the best strategy to maximize the metabolic energy for a neuron. Following on these ideas, here we examined if surprise minimization by single neurons could be a basis for consciousness. First, we showed in simulations that as a neural network learns a new task, then the surprise within neurons (defined as the difference between actual and expected activity) changes similarly to the consciousness of skills in humans. Moreover, implementing adaptation of neuronal activity to minimize surprise at fast time scales (tens of milliseconds) resulted in improved network performance. This improvement is likely because adapting activity based on the internal predictive model allows each neuron to make a more "educated" response to stimuli. Based on those results, we propose that the neuronal predictive adaptation to minimize surprise could be a basic building block of conscious processing. Such adaptation allows neurons to exchange information about own predictions and thus to build more complex predictive models. To be precise, we provide an equation to quantify consciousness as the amount of surprise minus the size of the adaptation error. Since neuronal adaptation can be studied experimentally, this can allow testing directly our hypothesis. Specifically, we postulate that any substance affecting neuronal adaptation will also affect consciousness. Interestingly, our predictive adaptation hypothesis is consistent with multiple ideas presented previously in diverse theories of consciousness, such as global workspace theory, integrated information, attention schema theory, and predictive processing framework. In summary, we present a theoretical, computational, and experimental support for the hypothesis that neuronal adaptation is a possible biological mechanism of conscious processing, and we discuss how this could provide a step toward a unified theory of consciousness.

Keywords: brain-inspired artificial neuronal networks, neuronal adaptation, theory of consciousness, biological learning algorithms, anesthesia

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Robert Kozma,
University of Memphis, United States
Steven Sloman,
Brown University, United States

****Correspondence:***
Artur Luczak
Luczak@uleth.ca

Received: *30 August 2021*
Accepted: *29 November 2021*
Published: *11 January 2022*

Citation:
Luczak A and Kubo Y (2022) Predictive Neuronal Adaptation as a Basis for Consciousness. Front. Syst. Neurosci. 15:767461. doi: 10.3389/fnsys.2021.767461

INTRODUCTION

"How does the brain work? Gather enough philosophers, psychologists, and neuroscientists together (ideally with a few mathematicians and clinicians added to the mix), and I guarantee that a group will rapidly form to advocate for one answer in particular: that the brain is a prediction machine" (Seth, 2020). Predictive processing was also suggested to be one of the most promising approaches to understand consciousness (Yufik and Friston, 2016; Hohwy and Seth, 2020). Nevertheless, it is still unclear how predictive processing could be implemented in the brain (Lillicrap et al., 2020), as most of the proposed algorithms require a precise network configuration (Rao and Ballard, 2005; Bastos et al., 2012; Whittington and Bogacz, 2017), which could be difficult to achieve, considering variability in neuronal circuits (y Cajal, 1911).

To address this problem, we proposed that single neurons can internally calculate predictions, which eliminates requirement of precise neuronal circuits (Luczak et al., 2022). Biological neurons have a variety of intracellular processes suitable for implementing predictions (Gutfreund et al., 1995; Stuart and Sakmann, 1995; Koch et al., 1996; Larkum et al., 1999; Ha and Cheong, 2017). The most likely candidate for realizing predictive neuronal mechanism appears to be calcium signaling (Bittner et al., 2017). For instance, when a neuron is activated, it leads to a higher level of somatic calcium lasting for tens of ms (Ali and Kwan, 2019). As neuron activity is correlated with its past activity within tens of ms (Harris et al., 2003; Luczak et al., 2004), thus, lasting increase in calcium concentration may serve as a simple predictive signal that a higher level of follow up activity is expected. Notably, basic properties of neurons are highly conserved throughout evolution (Kandel et al., 2000; Gomez et al., 2001; Roberts and Glanzman, 2003), therefore a single neuron with a predictive mechanism could provide an elementary unit to build predictive brains for diverse groups of animals.

This idea is further supported by a theoretical derivation showing that the predictive learning rule provides an optimal strategy for maximizing metabolic energy of a neuron. The details of derivation are described in a study (Luczak et al., 2022) and a summary is depicted in **Figure 1**. Shortly, E_b represents energy received from blood vessels in the form of glucose and oxygen, which is a non-linear function of local neuronal population activity, including the considered neuron j activity (x_j) (Devor et al., 2003; Sokoloff, 2008). The E_{ele} represents the energy consumed by a neuron for electrical activity, which is mostly a function of the presynaptic activity (x_i) and respective synaptic weights (w_{ij}) (Harris et al., 2012). A neuron also consumes energy on housekeeping functions, which could be represented by a constant E_h. As described in a study (Luczak et al., 2022), this formulation shows that to maximize energy balance, a neuron has to minimize its electrical activity (be active as little as possible), but at the same time, it should maximize its impact on other neurons' activities to increase blood supply (be active as much as possible). Thus, weights must be adjusted to strike a balance between two opposing demands: maximizing the neuron's downstream impact and minimizing its own activity (cost). This energy objective of a cell could be paraphrased as the "*lazy neuron principle: maximum impact with minimum activity.*" We can calculate such required changes in synaptic weights (Δw) that will maximize neuron's energy (E_j) by using gradient ascent method [for derivation see **Supplementary Material** or (Luczak et al., 2022)]. As a result, we found that maximizing future energy balance by a neuron leads to a predictive learning rule, where a neuron adjusts its synaptic weights to minimize surprise [i.e., the difference between actual (x_j) and predicted activity ($\tilde{x}_j$)].

Interestingly, this derived learning rule was shown to be a generalization of Hebbian-based rules and other biologically inspired learning algorithms, such as predictive coding and temporal difference learning (Luczak et al., 2022). For example, when $\tilde{x}_j = 0$ in our predictive learning rule (i.e., when a neuron

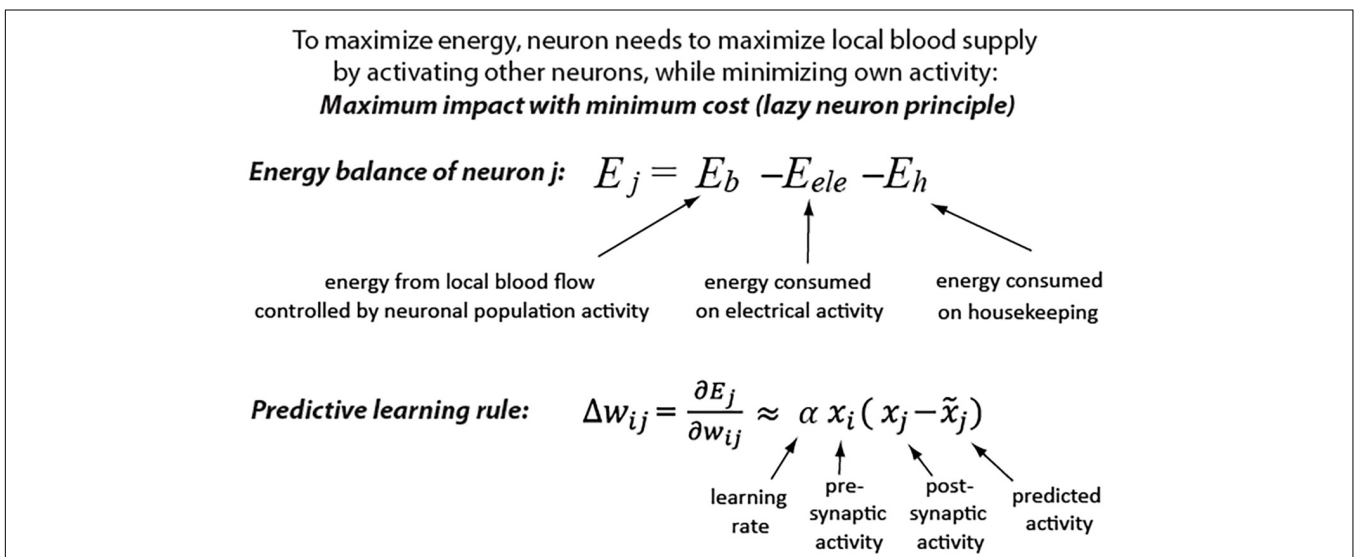


FIGURE 1 | Maximizing neuron metabolic energy leads to predictive synaptic learning rule (see **Supplementary Material** for derivation details).

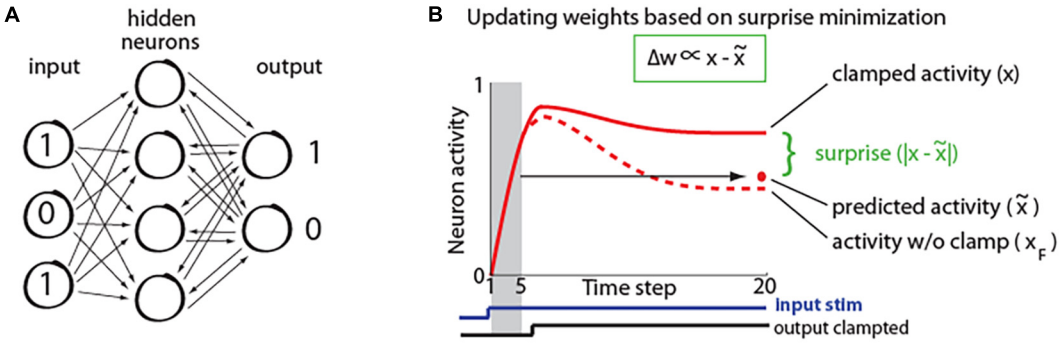


FIGURE 2 | (A) Simplified schematic of our recurrent network architecture. For visualization, only a small subset of neurons is shown. **(B)** Illustration of neuron activity in response to a stimulus. Initially the network receives only the input signal (bottom blue trace), but after 8 steps, the output signal is also presented (a.k.a. clamped phase; bottom black trace). The red dot represents steady-state activity which was predicted from initial activity (in shaded region). The dashed line shows activity of the same neuron in response to the same stimulus, if the output would not be clamped (x_F; a.k.a. free phase), which neuron "wants" to predict. Green insert: synaptic weights (w) are adjusted in proportion ($\propto$) to the difference between steady-state activity in clamped phase (x) and predicted activity ($\tilde{x}$) [adopted from Luczak et al. (2022)].

does not make any prediction), then we obtain Hebb's rule: $\Delta w_{ij} = \alpha x_i x_j$, a.k.a. "cells that fire together, wire together"(Hebb, 1949). Moreover, our model belongs to the category of energy-based models, for which it was shown that synaptic update rules are consistent with spike-timing-dependent plasticity (Bengio et al., 2017). Thus, this predictive learning rule may provide a theoretical connection between multiple brain-inspired algorithms and may offer a step toward development of a unified theory of neuronal learning.

The goal of this paper is to show that the properties ascribed to consciousness could be explained in terms of predictive learning within single neurons. For that, first, we will implement a predictive learning rule in an artificial neural network, and then we will use those simulation results together with biological evidence to propose a predictive neuronal adaptation theory of consciousness.

METHODS

Implementation of a Predictive Learning Rule in a Neural Network

To study how properties of predictive learning rule may relate to consciousness processes, we created a recurrent neural network. It had 420 input units, 50 hidden units, and 10 output units as illustrated in **Figure 2A**. The network was trained on a hand-written digit recognition task MNIST (LeCun et al., 1998), with 21 × 20 pixels from center of each image given as input to the network. The details of network training are described in a study (Luczak et al., 2022). First, network is presented with only an input signal and the activity starts propagating throughout the network until it converges to a steady-state, when the neurons' activity stops changing, as depicted in **Figure 2B**. This is repeated for 1,600 randomly chosen stimuli. During this phase, we also trained a linear model to predict the steady-state activity. Specifically, for each individual neuron, the activity during the five initial time steps $(x_{(1)}, \ldots, x_{(5)})$ was used to predict its steady-state activity at time step 20: $x_{(20)}$, such that: $x_{(20)} \approx \tilde{x} = \lambda_{(1)} * x_{(1)}, + \cdots + \lambda_{(5)} * x_{(5)} + b$, where $\tilde{x}$ denotes predicted activity, λ and b correspond to coefficients and offset terms of the least-squared model, and the terms in brackets correspond to time steps (**Figure 2B**). Next, a new set of 400 stimuli was used, where from step 8, the network output was clamped at values corresponding to image class (teaching signal). For example, if the image of number 5 was presented, then the value of the 5th output neuron was set to "1," and the values of the other 9 output neurons was set to "0," and network was allowed to settle to the steady-state. This steady-state was then compared with predicted steady-state activity, which was calculated using the above least-squared model. Subsequently, for each neuron, the weights were updated based on the difference between the actual (x_j) and its predicted activity $(\tilde{x}_j)$ in proportion to each input contribution (x_i), as prescribed by the predictive learning rule in **Figure 1** (Matlab code for a sample network with our predictive learning rule is provided in **Supplementary Material**).

RESULTS

Neuronal Surprise Reproduces Stages of Skill Consciousness

The network using predictive learning rule showed a typical learning curve, with rapid improvement in performance in the first few training epochs, and with plateauing performance during later training epochs (**Figure 3A**). Notably, this shape of learning curve is also typical for skill-learning in humans, where, initially at the novice level, there are fast improvements, and it takes exponentially more time to improve skills at, for example, elite athlete level (Newell and Rosenbloom, 1981). However, what is new and interesting here, is how a surprise (i.e., the difference between actual and predicted activity) evolved during network training (**Figure 3B**) and how it compares to the stages of "skill consciousness," as explained below.

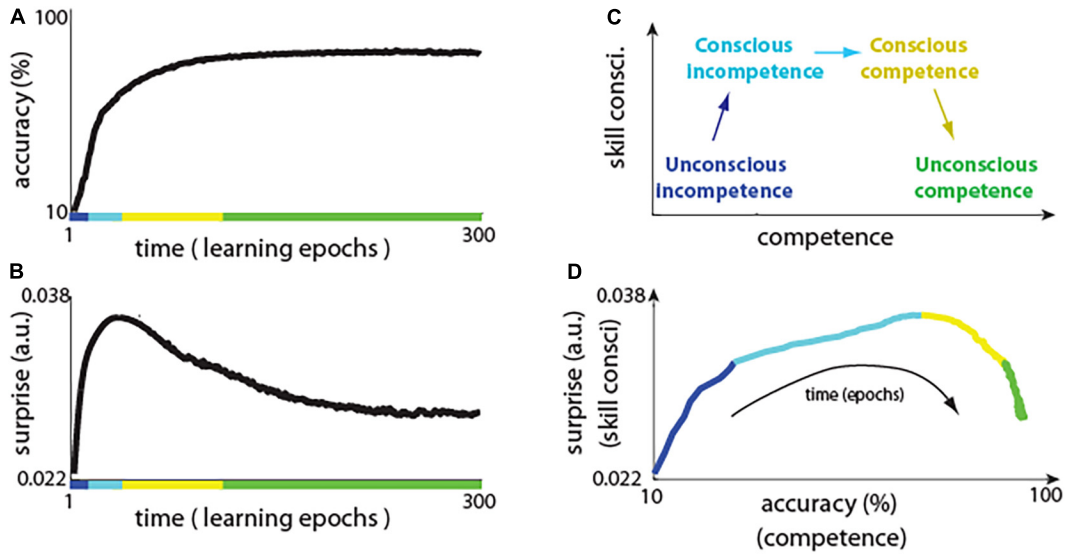


FIGURE 3 | Neuronal surprise in network with predictive learning rule. **(A)** Learning curve showing accuracy of the network across learning epochs. Colors indicate conceivable extents of stages of "conscious competence" shown in panel **(C)**. **(B)** Change in neuronal surprise (|actual – predicted activity|) averaged over all neurons (see main text for details). **(C)** Diagram of the four stages of "conscious competence" during skill learning. **(D)** Neuronal surprise (proxy of skill consci.) vs accuracy (a.k.a. competence) during network learning. This is replotting data from panels **(A)** vs **(B)**, where each point corresponds to a single time epoch. Note that although for example, "unconscious competence" (marked in green) spans over 150 epochs in panels **(A)** and **(B)**, in panel **(C)** those points are "compressed" as there is little change in accuracy, and in surprise during those epochs.

It was observed that learning involves the four stages of "conscious competence" (Broadwell, 1969; Das and Biswas, 2018; **Figure 3C**): (1) Unconscious incompetence – where individual does not know what he/she doesn't know, and, thus, that individual is not aware of his/her own knowledge deficiencies (e.g., foreigner may not know about certain local traffic regulations); (2) Conscious incompetence – where the individual recognizes his/her own lack of knowledge or skills but does not have those skills (e.g., a car passenger who does not know how to drive); (3) Conscious competence – where the individual develops skills but using it requires conscious effort (e.g., beginner car driver); (4) Unconscious competence - where due to extensive practice, the individual can perform learned tasks on "autopilot" (e.g., driving car on the same route every day).

Here we illustrate how the above stages of conscious competence could be recapitulated by the network with our predictive learning rule. We used the neuronal surprise as a proxy measure of consciousness, which is motivated by previous theoretical (Friston, 2018; Waade et al., 2020) and experimental work (Babiloni et al., 2006; Del Cul et al., 2007), which will be discussed in later sections. We calculated the surprise for each neuron j as: $< |x_j - \tilde{x}_j| >$, where $|\ldots|$ denotes absolute value, and $< \ldots >$ denotes average across all 400 images presented in a single training epoch. The neuronal surprise was defined as mean surprise across all of neurons. To better illustrate the network behavior, we also plotted accuracy (a.k.a. competence) vs surprise (a proxy of consciousness) (**Figure 3D**; model details and code to reproduce presented figures are included in **Supplementary Material**). Initially, when the network was presented with an input image, the neurons in the hidden layer could almost perfectly predict what will be the steady-state activity after the output units are clamped (**Figure 3B**, first few epochs). This is because the network starts with random connections and the signal coming from 10 output units is relatively week in comparison to the signal coming to the hidden layer from a much larger number of input units: 420. Thus, the steady-state activity, which neurons learn to predict when only the input image is presented, is not much different from the steady-state activity when input image is presented together with clamped outputs. This is like the "unconscious incompetence" stage, as the network is almost completely "not aware" of the teaching (clamped) signal (**Figure 3D**, blue line). However, as the activity of the output neurons is mostly correlated with any discrepancy between the actual and the predicted activity in hidden layer neurons, thus, the synaptic weights from output neurons are most strongly modified. Consequently, as the learning progresses, the hidden neurons are more and more affected by the output units, and their surprise: the discrepancy between actual and predicted activity, increases. This is analogous to the "conscious incompetence," where the network becomes "aware" of the clamped teaching signal, but the network has not yet learned how to classify images correctly (**Figure 3D**, light blue). In result, as magnitude of surprise $|x_j - \tilde{x}_j|$ increases, then other synaptic weights also start changing more, as prescribed by the predictive learning rule in **Figure 1**. Those synaptic updates made the activity driven by the input image, closer to the desired activity as represented by the clamped output units. This could be characterized as "conscious competence," where the surprise signal allows the network to learn and to become more competent on that task

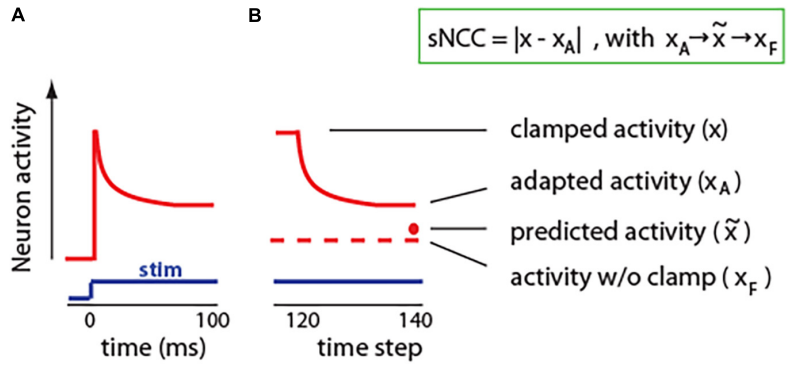


FIGURE 4 | Adaptation in neurons and in our model. **(A)** Cartoon illustration of typical neuronal adaptation in response to constant stimuli. Note that after the initial increase, neuronal activity moves closer toward the activity level without stimulus. **(B)** Similarly, in our model, adaptation shifts neuronal activity toward activity level without clamp (predicted activity). Insert: We propose to define a single-neuron correlate of consciousness (sNCC) as neuronal adaptation. Arrows denote that activity adapts toward predicted activity, which approximates activity without a clamp.

(**Figure 3D**, yellow line). Finally, as network learns to predict the image class with high accuracy, then the surprise (the difference between predicted and clamped teaching signal) is diminishing, which is analogous to an expert who achieved "unconscious competence" (**Figure 3D**, green line). This, that the neuronal surprise recapitulates the stages of conscious competence, by first increasing and then decreasing during learning, was a general phenomenon across different datasets and across diverse network architectures (**Supplementary Figure 1**).

Surprise Reduction by Neuronal Adaptation

Derivation of the predictive learning rule in **Figure 1** shows that the best strategy for a cell to maximize metabolic energy is by adjusting its synaptic weights to minimize surprise: $|x - \tilde{x}|$. However, this change in surprise does not need to take minutes or hours, as typically required for structural synaptic modification to occur (Xu et al., 2009). Neurons have adaptation mechanisms, which could serve to reduce surprise at a much faster time scale of tens of ms (Whitmire and Stanley, 2016).

Neural adaptation is a ubiquitous phenomenon that can be observed in neurons in the periphery, as well as in the central nervous system; in vertebrates, as well as in invertebrates (Whitmire and Stanley, 2016; Benda, 2021). Neuronal adaptation can be defined as the change in activity in response to the same stimulus. The stimulus can be a current injection into a single neuron or a sensory input like sound, light, or whisker stimulation. Usually, neuron activity adapts in exponential-like fashion, with rapid adaptation at the beginning, and then later plateauing at a steady-state value (**Figure 4A**). Typically, neuronal adaptation is shown as the decrease in activity in response to excitatory stimuli. However, neurons can also adapt by increasing its spiking ability when inhibitory stimulus is presented; for example, an injection of constant hyperpolarizing current (Aizenman and Linden, 1999). Thus, adaptation could be seen as change in neurons activity toward a typical or expected (predicted) level ($\tilde{x}$).

To investigate effect of adaptation on neuronal processing, we implemented a brain-inspired adaptation mechanism in the units in our network. For this, during the clamped phase from time step 8, the activity of each neuron was nudged toward the predicted state (**Figure 4B**). Specifically, the activity of neuron j at time step t was calculated as: $x_{j,t} = a * \tilde{x} + (1 - a) * \sum_i (w_{i,j} * x_{i,t-1})$, where $0 \leq a \leq 1$ is a parameter denoting strength of adaptation. For example, for $a = 0$, the adaptation is equal to zero, and the network activity is the same as in original network described in **Figure 2**. To update synaptic weights, we used the same learning rule as in **Figure 1**: $\Delta w_{i,j} = x_i(x_j - \tilde{x}_j)$, but here x_j represents clamped activity with added adaptation, which can also be denoted as x_A. Interestingly, networks with implemented adaptation achieved better accuracy than the same networks without adaptation (**Supplementary Information**). This could be due to the fact that if an activity in the clamped phase is much different from an expected activity without the clamp, then learning may deteriorate as those two network states could be in different modes of the energy function (Scellier and Bengio, 2017). Adaptation may reduce this problem by bringing clamped state closer to expected. To give an analogy, if part of a car is occluded by a tree, then, purely by sensory information, we cannot say what is behind that tree. However, based on our internal model of the world, we know what shape a car is, and, thus, we can assume that the rest of the car is likely behind the tree. Similarly, neuronal adaptation may allow a neuron to integrate the input information with predictions from its internal model, and then adjust its activity based on this combined information leading to a more appropriate response.

HYPOTHESIS AND THEORY

Predictive Adaptation as a Signature of Consciousness

It is largely accepted that consciousness is a gradual phenomenon (Francken et al., 2021). It was also suggested that even a single cell may have a minimum level of consciousness,

based on the complexity of behavior and complexity of information-processing within each cell (Reber, 2016; Baluška and Reber, 2019). For example, every single cell contains large biochemical networks, which were shown to make decisions and to perform computations comparable to electrical logic circuits (**Supplementary Figure 2**; McAdams and Shapiro, 1995). This allows for highly adaptive behavior, including sensing and navigating toward food, avoiding a variety of hazards, and coping with varying environments (Kaiser, 2013; Boisseau et al., 2016). For instance, single-celled organisms were shown to be able to "solve" mazes (Tero et al., 2010), to "memorize" the geometry of its swimming area (Kunita et al., 2016), and to learn to ignore irritating stimulus if the cell's response to it was ineffective (Tang and Marshall, 2018). Moreover, single-celled microorganisms were shown to predict environmental changes, and to appropriately adapt their behavior in advance (Tagkopoulos et al., 2008; Mitchell et al., 2009). Those complex adaptive behaviors were proposed to resemble cognitive behavior in more complex animals (Lyon, 2015). This likely requires organism to build some sort of internal predictive model of their own place in the environment, which could be considered as a basic requirement for consciousness.

The results presented in **Figure 3** suggest that the level of consciousness could be related to the amount of surprise. This is also supported by results from human EEG studies, where the neuronal signature of surprise: P300, closely reflects conscious perception (Del Cul et al., 2007; Dehaene and Changeux, 2011). Here, we propose that in a neuron, adaptation could be seen similarly to P300, as a measure of surprise, and thus, it could provide an estimate of the level of "conscious cellular perception." Specifically, as described above, surprise could be defined as a difference between actual (x) and predicted activity ($\tilde{x}$). Because adaptation changes neurons' activity toward a predicted activity level, thus, the size of adaptation ($|x - x_A|$) is directly related to the size of surprise: ($|x - \tilde{x}|$). Therefore, we propose to define the single-neuron correlate of consciousness (sNCC) as the magnitude of neuronal adaptation sNCC $= |x - x_A|$, (**Figure 4B**). Based on this, we hypothesize that single-cell predictive adaptation is a minimal and sufficient mechanism for conscious experience.

Generalized Definition of Consciousness as a Process of Surprise Minimization

First, we will explain the main ideas using a simplified example, then later, we will present how it can be generalized. Let us have a two-dimensional environment, where at each location P, there is a certain amount of food. There is also an organism that wants to go to a location with the highest amount of food. That organism does not know exactly how much food there is at any given location, but based on past experience, the organism has an internal model of the environment to help with predictions. For instance, let us assume that the maximum concentration of food *(m)* is at point P_m, but the smell of food comes from the direction of point: P_s, where s stands for sensory evidence (**Figure 5**). However, the concentration of food in the past was highest in the North direction. The internal model combines this information and predicts the highest probability of food in the North East direction at point P_p, where p stands for predicted. Based on this, the organism adapts and moves toward P_p to location P_A, where $_A$ stands for adaptation. When the organism arrives to P_A location, then it can compare the actual amount of food at that location with the predicted one, and update the internal model accordingly. Thus, by combining sensory information and internal model predictions, our organism was able to adapt its behavior more appropriately.

In the above-described case, we could say that our organism was quite conscious of its environment, as it made close to optimal decision. We can quantify it by measuring how close to optimal location an organism moved: $d(P_A, P_m)$, as compared if it would move in reflex-like fashion to location that is purely determined by sensory stimulus: $d(P_s, P_m)$, where d(.,.) denotes a distance between 2 points. Specifically, we can define organism consciousness of environment as $C_e = d(P_s, P_m) - d(P_A, P_m)$, (**Figure 5**, insert). It is worth noting that if an organism has a good model of external environment to correctly predict location with maximum food, then: $P_p \approx$

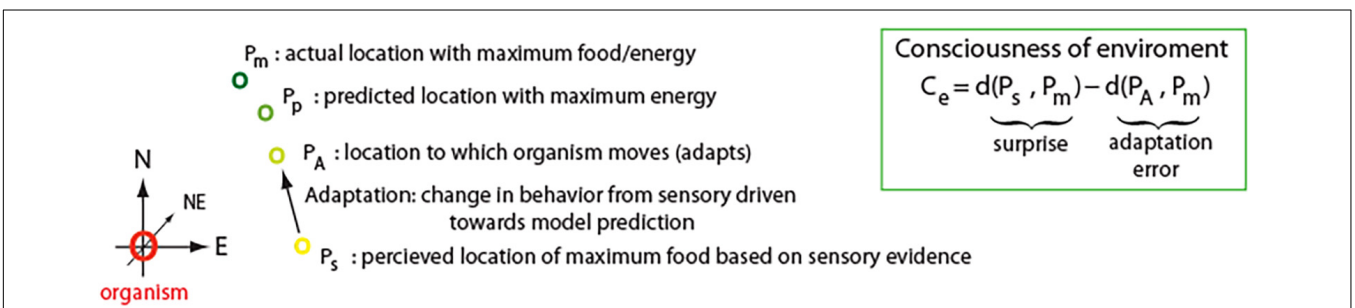


FIGURE 5 | Insert: Consciousness (C_e) is defined here as a surprise: distance d(,) between obtained sensory information (P_s) and expected information. However, if system cannot appropriately adapt based on that information, then conscious perception is reduced (adaptation error). Thus, Consciousness is a function of surprise and ability of organism to adapt to minimize that surprise. Expected information is denoted by P_p and it is calculated by internal predictive model, which based on partially available data tries to approximate actual state of the environment (P_m). Schematic on the left illustrates concept of C_e for sample organism living in 2D environment (see main text for details).

P_m, and thus, the first term in C_e : $d(P_s, P_m) \approx d(P_s, P_p)$, where this distance $d(P_s, P_p)$ between sensory evidence (P_s) and model prediction (P_p) is a description of surprise. The second term in C_e : $-d(P_A, P_m)$, describes how far organism is from location with maximum food/energy (P_m) after it made adaptation (P_A). This could be seen as an error term, which could arise if predictive model is incorrect or if organism is unable to move exactly to the predicted location. Hence, according to the above definition of C_e, consciousness is equivalent to surprise, if error term is 0, which would be the case for an organism to able to perfectly adapt.

Although, we used here a two-dimensional environment as an example, this can be generalized to a high-dimensional sensory space. Let us consider a simple organism which can sense concentration of 10 substances in a deep ocean. As organism swims, it changes its position in 3D space, but more importantly, concentration of 10 substances indicating location of food and predators also changed with each movement. Thus, 3D space translates to 10D sensory space, which is more relevant to that organism behavior. Therefore, distances d in $C_e = d(P_s, P_m) - d(P_A, P_m)$, may be more appropriately calculated in sensory space of that organism, instead of the standard 3D spatial coordinates. For example, we implicitly used idea of sensory space in case of neurons shown in **Figures 3**, **4**. Neuron senses its local environment through variety of channels located especially in synapses. Activity of a neuron affects other neurons, which through feedback loops change synaptic inputs to that neuron, and thus, its sensory environment. Because neuron gets energy from blood vessels, which dilation is controlled by coordinated activity of local neurons, therefore, neuron may "want" to move in the sensory space corresponding to activity patterns resulting in the most local blood flow. Therefore, change in neuron activity is equivalent to a movement in a chemical sensory space, where different locations in that space correspond to different amount of energy obtained by a neuron. For that, the word "environment" in C_e refers to this highly dimensional sensory space rather than that of the typical 3D space.

This generalization to sensory space also allows to see notions introduced earlier in **Figures 3**, **4** as special cases of environmental consciousness C_e. For example, when organism has the perfect model of external environment, then it can correctly predict the location with maximum food, thus, $P_p = P_m$, as we have explained before. However, if that organism can also move exactly to predicted location such that: $P_A = P_p$, then, also $P_A = P_m$. In such case, an adaptation error $d(P_A, P_m)$ becomes 0, and thus, $C_e = d(P_s, P_m)$. Considering the above case that $P_m = P_A$, C_e can also be expressed as $C_e = d(P_s, P_A)$, which is a distance by how much an organism moved or adapted. Thus, in case of the neuron described in **Figure 4**, $C_e = d(P_s, P_A) \approx d|x, x_A| = |x - x_A| =$ sNCC. Similarly, as mentioned earlier, C_e becomes equivalent to surprise if organism perfectly adapts ($P_A = P_p = P_m$). In such case, adaptation error is zero, and we can write $C_e = d(P_s, P_p) \approx |x - \tilde{x}|$, which is the distance between the stimulus-evoked activity and the model prediction, which we used to quantify the skill consciousness in **Figure 3**. Thus, C_e is a function of surprise and ability of organism to adapt to minimize that surprise.

Note that surprise and adaptation could be considered as contributing to C_e on different timescales, with synaptic changes gradually minimizing surprise over a long period of time, and with neuronal adaptation changing neuronal firing rapidly within 10–100 ms. When an organism is learning a new skill, then activity driven by bottom-up signals is different from activity provided by top-down teaching signals, which results in a higher surprise term. However, if neurons cannot adapt their activity accordingly (e.g., when biochemical processes mediating adaptation within a neuron are blocked), then adaptation error will be as large as the surprise term, resulting in $C_e = 0$ and, thus, in no conscious experience. Therefore, the surprise term could be interpreted as "potential consciousness," meaning the maximum possible consciousness to a given stimulus. Synaptic strength gradually changes over a period of learning, resulting in slow changes in "potential consciousness." However, when a stimulus is presented, and neurons rapidly adapt their activity toward the predicted level, it reduces the adaptation error term and results in $C_e > 0$, and, thus, in conscious perception within a fraction of a second.

Hypothesis Validation

A hypothesis, by definition, should generate testable predictions. Our main hypothesis is that the *neuronal adaptation is a neuronal correlate of consciousness*. This implies that neurons and, thus, brains, without adaptation cannot be conscious. Therefore, our hypothesis predicts that any mechanism which affects neuronal adaptation will also affect consciousness. This prediction was shown to be correct for a diverse group of neurochemicals involved in sleep and anesthesia, which also alter the neuronal adaptation. For instance, levels of multiple neuromodulators in the brain such as serotonin, noradrenaline, and acetylcholine are significantly different between waking and sleeping in REM or non-REM stages (España and Scammell, 2011). Whole-cell voltage-clamp recordings *in vitro* in the pyramidal neurons have demonstrated that all those neuromodulators also affect neuronal adaptation (Satake et al., 2008). Similar results were obtained when testing various substances used for anesthesia, such as urethane (Sceniak and MacIver, 2006), pentobarbital (Wehr and Zador, 2005), and ketamine (Rennaker et al., 2007). Moreover, it was shown that a large variety of anesthetics, including butanol, ethanol, ketamine, lidocaine, and methohexital are blocking calcium-activated potassium channels, which mediate neuronal adaptation (Dreixler et al., 2000). Interestingly, considering a broad spectrum of molecular and cellular mechanisms affected by different anesthetic compounds, there remains significant uncertainty of what is the single mechanism underlying anesthesia (Armstrong et al., 2018). Our theory suggests that what all anesthetics could have in common is the ability to disturb neuronal adaptation. Thus, our theory clearly provides testable predictions, which could either be invalidated or validated by using pharmacological and electrophysiological methods (see also "Limitation" section for more discussion on this topic).

Predictive Adaptation as a Step Toward a Unified Theory of Consciousness

Important consequence of a neuron adapting its activity toward a predicted level is that it allows neurons to exchange information about its predictions. Thus, neuron output activity is not exclusively driven by its synaptic inputs, but it is also a function of its internal predictive model. Below, we will briefly describe a few of the most prominent studies, as well as the theories of consciousness [for in-depth reviews see Francken et al. (2021) and Seth (2021)]. We will particularly focus on outlining the differences and similarities to our theory of predictive adaptation, and how it may provide a theoretical basis for connecting diverse theories of consciousness.

Connection to Optical Illusions

Exchanging predictions among neurons may explain multiple phenomena linked to conscious perception, such as optical illusions. For example, let us consider a neuron tuned to detect horizontal lines. Such neuron may learn that even when feed-forward inputs are not exactly consistent with a line (e.g., due to partial occlusion), then later on, it usually receives a top-down signal indicating detection of a line due to combining information from other parts of the image by higher cortical areas. Thus, in the case of an image with illusory contours, this neuron may receive less activation from feed-forward inputs, as parts of the lines are missing. However, based on experience with occluded objects, that neuron may predict that it will soon receive top-down signals indicating a line, thus, in expectation it will increase its activity toward predicted levels. Consequently, other neurons receiving this predictive information are more likely to interpret it as a line, resulting in positive feedback loops and coherent perception of a line.

Similar explanation could also be applied in case of ambiguous images like the Rubin vase (face) optical illusion. If a set of neurons in the association cortex receives inputs suggesting an image of a face, then they will increase their activity accordingly toward that "believe," triggering a global activity pattern giving a single perception of a face.

Connection to Global Neuronal Workspace Theory

As described above, a large-scale neuronal convergence to a single "believe" is very similar to a theory of global neuronal workspace (GNW) (Baars, 2002; Dehaene and Changeux, 2011). Briefly, GNW states that an organism is conscious of something, only when many different parts of the brain have access to that information. Additionally, if that information is contained only in the specific sensory or motor area, then the organism is unconscious of that something. In our theory, consciousness is on a continuous scale. However, if an activity is different from what is expected across the many parts of the brain, then our measure of C_e will also be larger as compared to a single brain area, and because the brain is a highly non-linear system, C_e could be orders of magnitude larger when the difference between expected and predicted signal is exchanged in feedback loops across the entire brain. Thus, if the brain during waking has close to maximum C_e, and low C_e during, for example, sleep, with intermediate values of C_e existing shortly during transition between those states, then this could reconcile the apparent difference between both theories. It is worth noting also that according to the GNW theory, a key signature of information accessing consciousness is the P300 component, which as mentioned earlier reflects surprise (Donchin, 1981; Mars et al., 2008). This is similar to our theory where C_e is defined in terms of surprise (**Figure 5**). Therefore, taken all together, GNW may be seen as a special case in our theory, where C_e is discretized to have only two values.

Connection to Integrated Information Theory

Our theory is also consistent with the main ideas of integrated information theory (IIT). The IIT quantifies consciousness as the amount of information generated by an integrated complex of elements above and beyond the information generated by its individual parts, which is denoted as Phi (Tononi, 2015; Tononi and Koch, 2015). Similarly, in our case, if two cells can communicate, then this will allow each of them to make better predictions and, thus, to increase combined C_e, by reducing error term: d (P_A, Pm). For instance, if cell #1, just by chance, has more receptors to detect substance *s1*, and cell #2 has slightly more receptors for substance *s2*, then by communicating predictions to each other, both cells will be able to better detect food, which secretes *s1* and *s2* [i.e., the wisdom of crowds (Friedman et al., 2001)]. This simplified example can be directly extended to neurons, where each has unique pattern of connections, thus, partly providing novel information to other neurons. However, there is one important difference between Phi and C_e. While Phi can be computed based purely on connectivity pattern, C_e also depends on stimulus. If stimulus is unexpected, then surprise term d (P_s, Pm) will increase, and thus, even without any change in network architecture, organism will be more conscious of that stimulus. However, on average, elements with more complex connectivity patterns, which have higher integrated information Phi, will also have higher C_e, as more information sources will be available for each element to improve predictions, thus, reducing error term in C_e.

Connection to Attention Schema Theory

It was also proposed that consciousness requires building an internal model of incoming information. For example, the brain constructs a simplified model of the body to help monitor and control movements, and similarly, at more abstract level, it may construct an internal model of attention, which could form a basis for consciousness (Graziano and Kastner, 2011; Graziano and Webb, 2015). In our theory, an internal model is a crucial part of defining the consciousness. Although our predictive model is at the single-cell level, communication between neurons could allow to form more complex models at the network level. Note that due to neuronal adaptation toward predicted activity, each neuron sends information to others, reflecting its internal model predictions. Thus, neurons in higher areas build their internal models based on combining information from other neuron models. This suggests that higher-order models, like the model of attention proposed by Graziano, could be a direct consequence

of building the brain from elements with internal models as described by our theory.

Connection to Predictive Processing

Our theory is closely related to the predictive processing framework. This theoretical framework posits that the brain's overall function is to minimize the long-term average prediction error (Hohwy and Seth, 2020). It also proposes that to accomplish this process, the brain needs to have a generative model of its internal and external environment, and continually update this model based on prediction error (Friston, 2005, 2010; Friston et al., 2017). The precursor of the predictive processing idea could be traced back to a 19th century scholar named Hermann von Helmholtz (Von Helmholtz, 1867). He suggested that the brain fills in some missing information to make a better sense of its surrounding environment. As in the earlier example of a car behind a tree, the brain fills in the occluded parts to provide the most likely picture of the surrounding world. Over the recent years, predictive processing has gained significant experimental support [see for review Walsh et al. (2020)]. There were also proposed predictive computational models of vision, illustrating how top-down processing can enhance bottom-up information (McClelland and Rumelhart, 1981; Rao and Ballard, 2005). An important theoretical advancement was made, when it was shown that predictive processing can be understood as Bayesian inference to infer the causes of sensory signals (Friston, 2003, 2005). This provided a mathematically precise characterization of the predictive processing framework, which was further generalized in the form of the free energy principle (Friston, 2010). Our theory is fully consistent with this framework. However, our work provides three novel and important contributions to predictive processing:

(1) We derived mathematically that the predictive processing maximizes metabolic energy of a neuron (**Figure 1**), which provides biologically bound theoretical basis for predictive processing framework.
(2) Based on the above theoretical considerations and based on computational simulations, we showed that a single neuron could be the basic element for building diverse predictive networks (**Figures 2**, **3**). This offers a solution to how predictive processing could be implemented in the brain without the need for precisely designed neuronal circuits or special "error units."
(3) Most importantly, we showed that predictive neuronal adaptation could be the mechanism for conscious processing (**Figure 4**) and based on this, we proposed a quantitative definition of consciousness (**Figure 5**).

LIMITATIONS

While the present study offers a novel theoretical model of consciousness derived from basic principles of maximizing metabolic energy, this also comes with caveats that should be considered. In the absence of a generally accepted definition and measure of consciousness, all theories of consciousness, including ours, are unfortunately more speculative than typical theories in mostly other areas of science. For instance, to date, no theory has convincingly demonstrated yet how neuronal mechanisms can generate a specific conscious experience. Similarly, with our theory, it has yet to be shown that connecting billions of adaptive neurons could result in subjective feelings of "self," which is typically considered as consciousness. Here, as a step toward addressing this problem, we described how single-neuron-level predictive processes could be related to consciousness of skills at the organism level (**Figure 3**). However, the caveat here is that "skills consciousness" (as well as "consciousness") does not have a well-defined measure, thus, changes in skill consciousness during learning are only described in loose qualitative terms. This needs to be more rigorously measured in the future to allow for more quantitative comparison to our model.

The related problem in theories of consciousness is the difficulty in proving causal mechanisms of consciousness. For example, in our definition of consciousness, the first term represents "surprise" (**Figure 5**), and as we described earlier, there is strong a experimental evidence relating surprise (e.g., P300) to conscious perception in humans. However, the caveat is that it is also possible that surprise could be correlated with consciousness without causing it, thus, experiencing surprise and acting on it may not be sufficient to create consciousness. Similarly, we described experimental evidence showing that a diverse group of neurochemicals involved in sleep and anesthesia also affects neuronal adaptation. However, this is also only a correlation, and to prove that neuronal adaptation causes consciousness, experiments controlling multiple confounding factors, and selective blockage of adaptation would be needed to provide a more conclusive answer.

One interesting feature of our definition of consciousness is its simplicity and scalability: the same simple equation can describe consciousness at the single-cell level as well as at the whole organism level. However, this could be taken as an argument against our theory, as the claim of consciousness in the single cell or in a robot could be considered as a "far cry" from the typically understood notion of consciousness. This is a valid objection. To address this semantic problem, we introduced a broader term, "consciousness of environment" (C_e; **Figure 5**). What we are proposing in this manuscript is that the consciousness of environment is on a continuous scale, and the consciousness that we are experiencing as humans is just an extreme case of the same process. To give an analogy, the celestial movement of planets was considered to be governed by different laws than earthly objects, but now we understand that the same gravitational laws could be used to describe the movement of objects at both scales, which we suggest could be similar with consciousness. Unfortunately, we are still missing experimental means to precisely measure consciousness, which makes theories of consciousness more difficult to verify, and thus, more speculative.

Moreover, surprise minimalization could also be achieved by other means than the intracellular predictive mechanism proposed here. For instance, multiple predictive coding networks have been developed, with specially designed neuronal circuits including "error units," which allow for comparing expected and actual activity (Rao and Ballard, 2005; Bastos et al., 2012; Whittington and Bogacz, 2017; Sacramento et al., 2018). Such networks can be trained using other biological learning rules, like spike-time-depended plasticity [STDP; (Bi and Poo, 2001)]

or some variation of Hebbian learning [e.g., BCM (Bienenstock et al., 1982)]. Thus, it is possible that consciousness in neuronal system may be created by predictive mechanisms implemented only at the network level. One problem with predictive coding only at the circuit level is that it requires precise connectivity, which could be difficult to achieve, considering the complexity and variability of neuronal dendritic trees. Here, deriving from the basic principle of metabolic energy maximization, we suggest that predictive neurons could provide an elementary unit from which a variety of predictive circuits could be built, thus solving the above implementation problem. Therefore, in addition to intracellular predictions, neurons may form predictive circuits, giving rise to enhanced predictive abilities that increase the level of consciousness in an animal, as discussed above in relation to attention schema theory. Those network-level interactions may lead to a rapid and exponential-like increase in C_e. However, contrary to many other theories of consciousness, we suggest that this increase in Ce will not result in qualitative change, and that consciousness from single-celled organisms to humans could be described on a continuous scale, as the same adaptive process of surprise minimization.

DATA AVAILABILITY STATEMENT

The datasets and our code is freely available online: https://people.uleth.ca/~luczak/PredC/. The code to reproduce results of this study is also provided in **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AL conceived the project, analyzed data, performed computer simulations, and wrote the manuscript. YK performed computer simulations and contributed to writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Compute Canada, NSERC, and CIHR grants to AL.

ACKNOWLEDGMENTS

We thank Edgar Bermudez-Contreras and Eric Chalmers for their helpful comments. We would also like to thank the reviewers for helping us to significantly improve this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnsys.2021.767461/full#supplementary-material

REFERENCES

Aizenman, C. D., and Linden, D. J. (1999). Regulation of the rebound depolarization and spontaneous firing patterns of deep nuclear neurons in slices of rat cerebellum. *J. Neurophysiol.* 82, 1697–1709. doi: 10.1152/jn.1999.82.4.1697

Ali, F., and Kwan, A. C. (2019). Interpreting in vivo calcium signals from neuronal cell bodies, axons, and dendrites: a review. *Neurophotonics* 7:011402. doi: 10.1117/1.NPh.7.1.011402

Armstrong, R., Riaz, S., Hasan, S., Iqbal, F., Rice, T., and Syed, N. (2018). Mechanisms of anesthetic action and neurotoxicity: lessons from molluscs. *Front. Physiol.* 8:1138. doi: 10.3389/fphys.2017.01138

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends Cogn. Sci.* 6, 47–52. doi: 10.1016/s1364-6613(00)01819-2

Babiloni, C., Vecchio, F., Miriello, M., Romani, G. L., and Rossini, P. M. (2006). Visuo-spatial consciousness and parieto-occipital areas: a high-resolution EEG study. *Cereb. Cortex* 16, 37–46. doi: 10.1093/cercor/bhi082

Baluška, F., and Reber, A. (2019). Sentience and consciousness in single cells: how the first minds emerged in unicellular species. *BioEssays* 41:1800229. doi: 10.1002/bies.201800229

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Benda, J. (2021). Neural adaptation. *Curr. Biol.* 31, R110–R116.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., and Wu, Y. (2017). STDP-compatible approximation of backpropagation in an energy-based model. *Neural Comput.* 29, 555–577. doi: 10.1162/NECO_a_00934

Bi, G.-Q., and Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* 2, 32–48. doi: 10.1523/jneurosci.02-01-00032.1982

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., and Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* 357, 1033–1036. doi: 10.1126/science.aan3846

Boisseau, R. P., Vogel, D., and Dussutour, A. (2016). Habituation in non-neural organisms: evidence from slime moulds. *Proc. R. Soc. B Biol. Sci.* 283:20160446. doi: 10.1098/rspb.2016.0446

Broadwell, M. M. (1969). Teaching for learning (XVI). *Gospel Guardian* 20, 1–3.

Das, M. S., and Biswas, D. (2018). Competence learning model. *Int. J. Manag. Technol. Eng.* 8, 1955–1964.

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Del Cul, A., Baillet, S., and Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol.* 5:e050260. doi: 10.1371/journal.pbio.0050260

Devor, A., Dunn, A. K., Andermann, M. L., Ulbert, I., Boas, D. A., and Dale, A. M. (2003). Coupling of total hemoglobin concentration, oxygenation, and neural activity in rat somatosensory cortex. *Neuron* 39, 353–359. doi: 10.1016/s0896-6273(03)00403-3

Donchin, E. (1981). Surprise!…surprise? *Psychophysiology* 18, 493–513.

Dreixler, J. C., Jenkins, A., Cao, Y.-J., Roizen, J. D., and Houamed, K. M. (2000). Patch-clamp analysis of anesthetic interactions with recombinant SK2 subtype neuronal calcium-activated potassium channels. *Anesthesia Analgesia* 90, 727–732. doi: 10.1097/00000539-200003000-00040

España, R. A., and Scammell, T. E. (2011). Sleep neurobiology from a clinical perspective. *Sleep* 34, 845–858. doi: 10.5665/SLEEP.1112

Francken, J., Beerendonk, L., Molenaar, D., Fahrenfort, J., Kiverstein, J., Seth, A., et al. (2021). An academic survey on theoretical foundations, common assumptions and the current state of the field of consciousness science. *PsyArxiv* [Preprint]. doi: 10.31234/osf.io/8mbsk

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. New York, NY: Springer.

Friston, K. (2003). Learning and inference in the brain. *Neural Netw.* 16, 1325–1352. doi: 10.1016/j.neunet.2003.06.005

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2018). Am I self-conscious?(Or does self-organization entail self-consciousness?). *Front. Psychol.* 9:579. doi: 10.3389/fpsyg.2018.00579

Friston, K. J., Parr, T., and de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Gomez, M., De Castro, E., Guarin, E., Sasakura, H., Kuhara, A., Mori, I., et al. (2001). Ca2+ signaling via the neuronal calcium sensor-1 regulates associative learning and memory in *C. elegans*. *Neuron* 30, 241–248. doi: 10.1016/s0896-6273(01)00276-8

Graziano, M. S., and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cogn. Neurosci.* 2, 98–113. doi: 10.1080/17588928.2011.565121

Graziano, M. S., and Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol.* 6:500. doi: 10.3389/fpsyg.2015.00500

Gutfreund, Y., Yarom, Y., and Segev, I. (1995). Subthreshold oscillations and resonant-frequency in guinea-pig cortical-neurons - physiology and modeling. *J. Physiol. London* 483, 621–640. doi: 10.1113/jphysiol.1995.sp020611

Ha, G. E., and Cheong, E. (2017). Spike frequency adaptation in neurons of the central nervous system. *Exp. Neurobiol.* 26, 179–185. doi: 10.5607/en.2017.26.4.179

Harris, J. J., Jolivet, R., and Attwell, D. (2012). Synaptic energy use and supply. *Neuron* 75, 762–777. doi: 10.1016/j.neuron.2012.08.019

Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G., and Buzsáki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature* 424, 552–556. doi: 10.1038/nature01834

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Hohwy, J., and Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* 1:64.

Kaiser, A. D. (2013). Are myxobacteria intelligent? *Front. Microbiol.* 4:335. doi: 10.3389/fmicb.2013.00335

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*. New York, NY: McGraw-Hill.

Koch, C., Rapp, M., and Segev, I. (1996). A brief history of time (constants). *Cereb. Cortex* 6, 93–101. doi: 10.1093/cercor/6.2.93

Kunita, I., Yamaguchi, T., Tero, A., Akiyama, M., Kuroda, S., and Nakagaki, T. (2016). A ciliate memorizes the geometry of a swimming arena. *J. R. Soc. Interface* 13:20160155. doi: 10.1098/rsif.2016.0155

Larkum, M. E., Zhu, J. J., and Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* 398, 338–341. doi: 10.1038/18686

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nat. Rev. Neurosci.* 21, 335–346. doi: 10.1038/s41583-020-0277-3

Luczak, A., Hackett, T. A., Kajikawa, Y., and Laubach, M. (2004). Multivariate receptive field mapping in marmoset auditory cortex. *J. Neurosci. Methods* 136, 77–85. doi: 10.1016/j.jneumeth.2003.12.019

Luczak, A., McNaughton, B. L., and Kubo, Y. (2022). Neurons learn by predicting future activity. *Nat. Mach. Intelligence* [Epub ahead of pint].

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Front. Microbiol.* 6:264. doi: 10.3389/fmicb.2015.00264

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., et al. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *J. Neurosci.* 28, 12539–12545. doi: 10.1523/JNEUROSCI.2925-08.2008

McAdams, H. H., and Shapiro, L. (1995). Circuit simulation of genetic networks. *Science* 269, 650–656. doi: 10.1126/science.7624793

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* 88:375. doi: 10.1037/0033-295x.88.5.375

Mitchell, A., Romano, G. H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., et al. (2009). Adaptive prediction of environmental changes by microorganisms. *Nature* 460, 220–224. doi: 10.1038/nature08112

Newell, A., and Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cogn. Skills Acquisition* 1, 1–55.

Rao, R. P., and Ballard, D. H. (2005). "Probabilistic models of attention based on iconic representations and predictive coding," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. K. Tsotsos (Amsterdam: Elsevier), 553–561. doi: 10.1016/b978-012375731-9/50095-1

Reber, A. S. (2016). Caterpillars, consciousness and the origins of mind. *Anim. Sentience* 1:1. doi: 10.1017/cbo9780511812484.003

Rennaker, R., Carey, H., Anderson, S., Sloan, A., and Kilgard, M. (2007). Anesthesia suppresses nonsynchronous responses to repetitive broadband stimuli. *Neuroscience* 145, 357–369. doi: 10.1016/j.neuroscience.2006.11.043

Roberts, A. C., and Glanzman, D. L. (2003). Learning in Aplysia: looking at synaptic plasticity from both sides. *Trends Neurosci.* 26, 662–670. doi: 10.1016/j.tins.2003.09.014

Sacramento, J., Costa, R. P., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. *Adv. Neural Inform. Process. Syst.* 31, 8721–8732.

Satake, T., Mitani, H., Nakagome, K., and Kaneko, K. (2008). Individual and additive effects of neuromodulators on the slow components of afterhyperpolarization currents in layer V pyramidal cells of the rat medial prefrontal cortex. *Brain Res.* 1229, 47–60. doi: 10.1016/j.brainres.2008.06.098

Scellier, B., and Bengio, Y. (2017). Equilibrium propagation: bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* 11:24. doi: 10.3389/fncom.2017.00024

Sceniak, M. P., and MacIver, M. B. (2006). Cellular actions of urethane on rat visual cortical neurons in vitro. *J. Neurophysiol.* 95, 3865–3874. doi: 10.1152/jn.01196.2005

Seth, A. (2020). "Preface: the brain as a prediction machine," in *The Philosophy and Science of Predictive Processing*, eds D. Mendonça, M. Curado, and S. S. Gouveia (London: Bloomsbury Academic).

Seth, A. (2021). *Being You: A New Science of Consciousness*. Westminster: Penguin.

Sokoloff, L. (2008). The physiological and biochemical bases of functional brain imaging. *Adv. Cogn. Neurodyn.* 2, 327–334. doi: 10.1007/s11571-007-9033-x

Stuart, G., and Sakmann, B. (1995). Amplification of EPSPs by axosomatic sodium channels in neocortical pyramidal neurons. *Neuron* 15, 1065–1076. doi: 10.1016/0896-6273(95)90095-0

Tagkopoulos, I., Liu, Y.-C., and Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. *Science* 320, 1313–1317. doi: 10.1126/science.1154456

Tang, S. K., and Marshall, W. F. (2018). Cell learning. *Curr. Biol.* 28, R1180–R1184.

Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D. P., Fricker, M. D., et al. (2010). Rules for biologically inspired adaptive network design. *Science* 327, 439–442. doi: 10.1126/science.1177894

Tononi, G. (2015). Integrated information theory. *Scholarpedia* 10:4164.

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140167.

Von Helmholtz, H. (1867). *Treatise on Physiological Optics*. New York, NY: Dover Publications.

Waade, P. T., Olesen, C. L., Ito, M. M., and Mathys, C. (2020). Consciousness fluctuates with surprise: an empirical pre-study for the synthesis of the free energy principle and integrated information theory. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/qjrcu

Walsh, K. S., McGovern, D. P., Clark, A., and O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a

model of perception. *Ann. N. Y. Acad. Sci.* 1464:242. doi: 10.1111/nyas.14321

Wehr, M., and Zador, A. M. (2005). Synaptic mechanisms of forward suppression in rat auditory cortex. *Neuron* 47, 437–445. doi: 10.1016/j.neuron.2005.06.009

Whitmire, C. J., and Stanley, G. B. (2016). Rapid sensory adaptation redux: a circuit perspective. *Neuron* 92, 298–315. doi: 10.1016/j.neuron.2016.09.046

Whittington, J. C., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.* 29, 1229–1262. doi: 10.1162/NECO_a_00949

Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., et al. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* 462, 915–919. doi: 10.1038/nature08389

y Cajal, S. R. (1911). *Histologie du Système Nerveux de L'homme & Des Vertébrés: Cervelet, Cerveau Moyen, Rétine, Couche Optique, Corps Strié, Écorce Cérébrale Générale & Régionale, Grand Sympathique: A. Maloine*. Paris: Maloine.

Yufik, Y. M., and Friston, K. (2016). Life and understanding: the origins of "understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Systems Neuroscience

PERSPECTIVE
published: 02 March 2022
doi: 10.3389/fnsys.2022.788395

Frontopolar Cortex Specializes for Manipulation of Structured Information

James Kroger[1]* *and Chobok Kim*[2]

[1] Department of Psychology, New Mexico State University, Las Cruces, NM, United States, [2] Department of Psychology, College of Social Sciences, Kyungpook National University, Daegu, South Korea

Keywords: integration, frontopolar cortex, frontal cortex, cognition, working memory, fMRI

INTRODUCTION

It is only in the last 20 years that frontopolar cortex (FPC) has been recognized as distinct anatomically and functionally from dorsolateral prefrontal cortex (DLPFC). It has appeared to be recruited for complex or abstract cognition, and as a result has been thought to be responsible for the most sophisticated human understanding (Thiebaut de Schotten et al., 2017, #27). In this perspective article, we review recent thinking about frontal lobe organization, evidence bringing it into question, and revisit an alternative view of FPC function. We then present an original study arising from that view that demonstrates a new specialization of FPC.

Recently, several researchers have proposed a caudal-rostral organization of function in the frontal lobes, with the most complex or abstract information processing found in the most anterior portion of frontal cortex (Krawczyk et al., 2011; Voytek et al., 2015; Nee and D Esposito, 2016; Dixon et al., 2017; Duverne and Koechlin, 2017; Badre and Nee, 2018; Jerath et al., 2019; Sarafyazd and Jazayeri, 2019; Eichenbaum et al., 2020; Riddle et al., 2020). The nature of the cognitive function employed in these studies has not been uniform. Badre and Nee (2018) reviewed relevant literature, and noted that the output of each level of abstraction may feed into the next-lower level as top-down control signals constraining processing in the lower level, which in turn feeds into a lower level. At the lowest level of abstraction, premotor cortex produces information about appropriate responses that are fed to the motor region. Recently, strong evidence of connected regions in the frontal cortex has been produced by examining connectivity patterns (Thiebaut de Schotten et al., 2017). These cortical areas may be part of cortico-striatal loops arranged hierarchically (Mestres-MissÈ et al., 2012; Korb et al., 2017; Rusu and Pennartz, 2019).

Badre and Nee (2018) noted different kinds of abstraction in different studies. Some employ temporal abstraction, in which more temporally distant information is treated as more abstract (such as long-term future plans) and more temporally immediate information (an immediate choice, such as which direction to take at an intersection) is more concrete. For example, Dixon et al. (2017) studied three neural networks, each extending across multiple brain structures and brain lobes. The network including the frontal pole processed distal goals, such as career choices, and decisions subserving that goal, while the other networks processed more immediate goals.

Badre and Nee (2018) also noted that kinds of abstraction varied across studies. Some consisted of what he called "policy abstraction:" the addition of more rules as the context in which problems were solved; "relational integration abstraction," in which more stimulus dimensions had to be integrated in making responses; "temporal abstraction," such that contexts were retained

Edited by:
Yan Mark Yufik, Virtual Structures Research Inc., United States

Reviewed by:
Artur Luczak, University of Lethbridge, Canada
Mark Latash, The Pennsylvania State University (PSU), United States

***Correspondence:**
James Kroger
jkroger@nmsu.edu

Received: *02 October 2021*
Accepted: *20 January 2022*
Published: *02 March 2022*

Citation:
Kroger J and Kim C (2022) Frontopolar Cortex Specializes for Manipulation of Structured Information. Front. Syst. Neurosci. 16:788395. doi: 10.3389/fnsys.2022.788395

over increasing time intervals; and "domain general abstraction," meaning that anterior regions dealt with more domain-general information than caudal areas. As Badre notes, studies have not included more than one type of abstraction, and it's difficult to determine whether these kinds of abstraction are arrayed across the frontal lobes in similar ways, though they have produced caudal-to-rostral patterns of activation with increased abstraction. This raises the question of whether there is some operation common to all of these types of abstraction, and perhaps other kinds as well.

In fact, Badre and Nee (2018) suggest that the frontal pole may not be at the apex of the frontal hierarchy. Connectivity in the frontal lobe (and the rest of the brain) has been extensively examined. It may be assumed that information flows "down" from the "top" of a hierarchy toward the lower levels. In the parlance of a frontal hierarchy from frontal pole to motor cortex, there is more need for connections traveling from frontal pole toward lower level structures than for connections in the opposite direction, so that there is asymmetry in connections between frontal pole and lower areas. However, frontal pole exhibits more symmetry with other frontal areas than this scheme suggests. Instead, it is dorsolateral prefrontal cortex (DLPFC), specifically Brodmann areas 45 and 46, that exhibit the asymmetry that should be characteristic of the apex of the hierarchy. From a different perspective, this conclusion is supported by a near-infrared study by Schumacher et al. (2019).

A recent extensive review by Mansouri et al. (2020) provides a complex and comprehensive analysis of literature and portrays the ability to use abstract cognition as having a multitude of subprocesses, located in regions across the frontal cortex. A common network in prefrontal cortex, premotor areas, and posterior parietal (mostly intraparietal sulcus) is augmented by cognitive skills that together manifest in many areas of the brain. This approach holds promise for parcellating and identifying the components of higher cognition and their neural substrates. This is largely in agreement with a review by Dixon et al. (2017), which identifies three networks comprised of regions across the major parts of the cerebrum that have unique domains and together accomplish complex processing. These reviews provide evidence that cognitive control for processing complex or abstract tasks in the service of goal attainment may not be simply rooted in frontopolar cortex.

This leaves us in something of a quandary. There is a long history of observations of activity in the frontal pole accompanying the most complex task performance, yet it may not be passing the results "down" to constrain processing at lower levels, until motor cortex executes some response. Mediating the most abstract processing may not equate with being at the top of a command structure for executing tasks that involve abstraction. The anatomy seems to support just as well the idea that the most complex or abstract processing demands are "handed off" to the frontal pole, which is able to resolve abstract demands and return the result to the executive in DLPFC, which then determines a response that is translated into action in DLPFC or premotor. It also supports a model in which the frontal pole does not act this independently, but rather augments or joins functionally with DLPFC, when complex or abstract tasks must be mediated, by virtue of the highly integrative structure of the neuropil there (Jacobs et al., 1997, 2001). What has evolution yielded by adding the frontal pole to the executive? Perhaps it is some computational ability that is not part of the executive control of action, or cognitive control. Kroger et al. (2008) found that as subjects formed mental models to solve very complex problems, frontal pole was recruited. These models involved a high degree of relational complexity, as the models were created under the constraints of the problem. Resolving problems that are relationally complex has been shown to recruit frontopolar cortex (Kalina Christoff et al., 2001; Kroger et al., 2002, 2004; Wendelken et al., 2008; Bunge et al., 2009; Crone et al., 2009; Krawczyk et al., 2011; Bazargani et al., 2014).

Clearly humans are capable of more intelligent and creative cognition than higher primates. In particular, they excel at producing problem solutions which incorporate information not present in the problem and not dependent on external constraints. Yufik (2019) has proposed a model of intelligence and understanding that depends on creation of mental models by the neural substrate, which directly addresses the creative production of novel information. In this view, model construction is decoupled from sensory-motor flow, a notion compatible with frontal pole working outside of and in support of executive control. Yufik's model proposes specific neuronal processes depending on "neuronal packets" underlying creative understanding. Yufik and Friston (2016) provide an extensive foundation for the model.

The idea of a cognitive control hierarchy flowing from frontal pole posteriorly so that concrete motor behavior can execute the actions dictated by the cognitive control architecture makes many assumptions about the nature of information processing in the frontal lobes. Working memory does not only hold behavioral demands or control information and a person is not always in the act of executing actions in the service of abstract goals. Nonetheless, recent understanding of the frontal lobes arises from studies limited to cognitive control in goal satisfaction. What has not been discussed is the nature of the neural and psychological processing that happens in these frontal hierarchy studies, regardless of the kind of abstraction involved. The computations in neural circuits are difficult to discern and may depend for progress on theoretical approaches such as that of Yufik (2019). In studies where subjects execute tasks continuously with any of the kinds of abstraction discussed above, at the instant a subject sees a stimulus, they must form an arbitrarily complex representation—whether it is composed of rules, dimensions, temporal character, or domain information—and make a judgment or response according to the instructions of the experiment, which are also incorporated into the formed representation. If this representation is complex, it is likely that some refreshing reinstantiates the representation for maintenance. In everyday life, such representations are made frequently. It is possible that in the course of reasoning or planning, such a representation must be manipulated or altered. When altered, a new representation results. It may have retained much of the structure of the previously held representation, with changes. Reasoning may then entail creating a series of

representations, each derived from the previous representations, with some degree of maintained structure.

A potential shortcoming of hierarchical theories is that they posit that frontopolar cortex is recruited in the course of cognitively abstract or complex mentation. Yet, the frontal pole is recruited in paradigms that would be difficult to classify as abstract or complex. Pollmann et al. (2000) presented subjects with stimuli that contained a field of squares in one color, sometimes with a single square having a different color. The entire field moved back and forth in sinusoidal motion, and sometimes, a single square moved in a sinusoidal direction different from the field's. So, one square differed in color or motion. Subjects performed search on a series of stimuli, during which the defining feature distinguishing the single square altered between color or motion dimensions on successive trials. Frontopolar activity was observed during such changes in target dimension. When the dimension changed, the subject had to quickly manipulate their representation of the task.

Sweeney et al. (1996) conducted an anti-saccade task, in which subjects focused on a fixation, and a stimulus appeared somewhere quickly and disappeared. In saccade trials, the subjects looked at the spot where the stimulus had appeared. On anti-saccade trials, they were to look at a spot opposite the location of the stimulus, relative to the fixation. On anti-saccade trials, frontopolar cortex was recruited. On anti-saccade trials, subjects were required to form a cognitively more complex representation of the task.

These paradigms don't involve abstract representations recruiting frontal pole as prescribed by hierarchical organization theories. Abstraction or complexity is often created by compounding contingencies; both of these tasks seem to involve a single, one-level contingency which must be modified. They do involve manipulating or changing their representation of the task.

One theme common to many studies of FPC is the integration of information, which we refer to as structured information. An integrative role is supported by anatomical features of FPC, which differs from DLPFC in several respects. Pyramidal neurons there are sparser but have richer, more complex dendritic trees which receive more inputs than other association cortex and their intracortical connections are primarily to other supramodal association cortex (Jacobs et al., 1997, 2001). This morphology suggests a role of integrating function or representations across the higher processing centers in the brain. It is the most recently evolved part of the frontal lobes (Semendeferi et al., 2001) and is a late cortical structure to reach maturation (Flechsig, 1901; Gogtay et al., 2004) which can be delayed by years in those with higher IQ (Shaw et al., 2006). Developmental trends in the ability to handle increased cognitive complexity are well documented (Andrews and Halford, 2002; Loewenstein and Gentner, 2005; Uttal et al., 2008) and correspond to the maturation of FPC and frontal cortex in general (Bunge et al., 2002; Segalowitz and Davies, 2004). Any complex representation or task set would be well supported by this architecture, as would coordination of multiple representations, tasks, or cognitive operations.

Prabhakaran et al. (2000) performed a study in which maintenance of an integrated representation recruited FPC, along with DLPFC. Study participants viewed multiple letters and multiple locations denoted by brackets "[]" arranged in a sample array. In one condition, the letters were located in the center of the display, and the locations were distributed around the display. After a delay, participants indicated whether the letters, or the locations, or both, in a probe matched those in the sample. In another condition, each letter was located within one pair of brackets, which were distributed around the screen and participants judged whether the letters in the probe were located in the same locations as in the sample. Thus, participants maintained integrated representations of the letters and positions during the delay, and FPC responded to this task demand, but not during the other conditions not requiring integration. Some reservation about this interpretation is possible, however, since the number of stimuli participants maintained approaches working memory capacity (Cowan, 2001). Rypma et al. (1999) found FPC to be recruited when participants simply maintained six, but not four, letters in a match-to-sample paradigm. Clearly overtaxing memory capacity, Grasby et al. (1994) showed an activation in this area when subjects heard fifteen words and had to immediately recall them but not for the same task using a word list of five. Christoff and Gabrieli (2000) suggested that this may be due to use of a mnemonic strategy employed when capacity is exceeded. It is also possible that participants prevented decay of items in working memory by continually refreshing them. Badre and Wagner (2004) and Johnson et al. (2005) observed FPC recruitment when participants refreshed items in memory. Thus FPC activation found by Prabhakaran et al. (2000) may also have arisen from executive control processes maintaining a large, integrated representation.

De Pisapia et al. (2007) illustrated FPC recruitment for integration in a different paradigm. They required a number and operation (e.g., 9+) to be integrated with a subsequently viewed subtask (3 × 7). When subjects performed this integration, FPC was activated, but not when the subtask was presented and completed first. The authors claim, "integration within WM occurs when the result of a subtask becomes combined with an already ongoing main task," and emphasize that "integration is not just insertion of WM contents into another representation, but also requires that insertion follows and depends upon subtask processing." (p. 933). In this account, linkage of items by a task context is a key demand. In another study, Reynolds et al. (2006) observed bilateral FPC activity when subjects judged whether each of two words was concrete or abstract, then indicated whether the outcomes of the two judgments were the same or different. The emphasis in this paradigm was on integration of internally-generated information: results of these internal judgments were compared in working memory for sameness. Reynolds et al. (2006) propose that FPC responded to integration of the two words in the comparison act. Beyond being integrated for the comparison, the integrated working memory contents were not ancillary to execution of a task. In both of these studies, integration of information was a dynamic process executed by the participant to compute a novel task solution.

FPC engagement by integration has been observed in other studies. Fangmeier et al. (2006) conducted a study of three-term reasoning in which participants viewed in sequence three problem parts such as (1) × g, (2) g m, and (3) × m, and

indicated whether the third relationship followed from the first two. Capturing the separate neural responses to presentation of each of the three parts, they observed that FPC was activated when the second part was presented. At that point it seems participants, anticipating the form of the problem and third part, integrated the first two parts into a unitary representation. Green et al. (2006) observed FPC activation when stimuli were evaluated for analogical relationship, requiring complex relational integration, but not when similarity in categorical or semantic relationships were judged. Kosslyn et al. (1994) required participants to judge whether a heard word and a seen picture matched, resulting in FPC activation as the two stimuli were integrated in comparison. Strange et al. (2001) found that when making categorical decisions about letter strings, FPC responded when the rule defining the category was changed, inducing attempts to understand the relationships in the strings described by the new rules. It's likely this entailed integrating representations of hypothetical relationships. In another study, items were judged for the presence of simple features or abstract features (Goldberg et al., 2007). The difference between the two kinds of features lies in whether they were perceptual in nature (simple) or could be derived by verbal description (abstract). FPC was recruited when assessing the presence of the abstract features, probably because the descriptive nature of the feature entailed integrating a complex propositional representation of the feature. Monti et al. (2007) also found FPC activation increased during solution of difficult deduction problems compared to simpler deduction problems. In these deduction tasks and other high-level tasks like the Tower of Hanoi or Ravens Progressive Matrices it is necessary for subjects to integrate together a complex configuration of problem elements, and this task element is one possible key to their FPC recruitment.

Some attempts to contrast manipulation and maintenance have examined working memory for verbal material in modified match-to-sample paradigms employing letters (D'Esposito et al., 1999) or words and non-words (Barde and Thompson-Schill, 2002). In both studies, subjects determined whether a probe item was included in the sample. In some trials, the judgment included determining what position the item occupied in the sample set. In the manipulation condition the letters or non-words were reordered into alphabetical order, and Barde and Thompson-Schill included an additional manipulation condition in which words were arranged according to the size of the objects they referenced. Barde and Thompson-Schill analyzed activity by region, and grouped FPC and DLPFC together. This ROI produced stronger activation during the manipulate conditions. D'Esposito et al. (1999) analyzed neural responses in individual subjects separately, subtracting activation for maintenance from manipulation activity; most of their six subjects exhibited greater activity in FPC during manipulation. Since Barde and Thompson-Schill employed alphabetization and size ordering the manipulation elicited by these tasks also involved retrieval from long term memory, for either knowledge about alphabetical order or semantic memory about size. Retrieval of semantic information from long-term memory has been associated with FPC in verb-generation tasks without manipulation (Petersen et al., 1988; MacLeod et al., 1998). More importantly for the present discussion, these studies resemble the self-ordered tasks of Petrides et al. (1993) in that a set of stimuli are progressively altered, requiring constant creation of a structured representation *via* processing. They are not designed to discriminate the contributions of integration and manipulation.

The distinction we make between neural demands of integrating of information and manipulation of information echoes previous theoretical discussion about frontal lobe operation. Wood and Grafman reviewed theories of frontal lobe function and distinguished them along a process vs. representation scheme (Wood and Grafman, 2003). Extending this distinction to frontopolar cortex, studies which have focused on the integration of information best correspond to a representational view of FPC function, while depicting FPC as executing or managing manipulation resembles process-oriented theories of frontal lobe function.

There has been no direct comparison of representing integrated information, where representation is the primary cognitive task, and manipulation of information, in which information processing is key. The first goal of the current study is to determine whether representing integrated information, in the absence of manipulation or a task execution context, depends on FPC. We employed a delayed match-to-sample paradigm in which three letters, of different colors and placed in different locations, are maintained in memory and compared to a probe (see **Figure 1**). Neural responses were compared to a control task in which three white letters centrally located were retained and compared to a similar probe. To compare brain processing during manipulation of internal representations and representation of integrated information, another condition required making one of two changes to the integrated representation of the sample in memory. After presentation of the sample, and before presentation of the probe, a cue screen appeared instructing participants to change the identity or position of one of the sample letters. Then, the modified representation was compared to the probe to assess match. In this way, we contrasted FPC recruitment during maintenance of an integrated representation with the manipulation of it.

MATERIALS AND METHODS

Participants

Fifteen right-handed healthy subjects (age: 18–34; five female) participated in the study. No subject had a history of neurological or psychiatric problems. All participants completed informed consents and the research was approved by the University of New Mexico Institutional Review Board.

Cognitive Tasks

Stimuli were presented using a program written in E-Prime[1] and back-projected onto a screen, sitting outside the magnet bore 37.5 inches from a mirror mounted over the participant's eyes

[1]www.pstnet.com

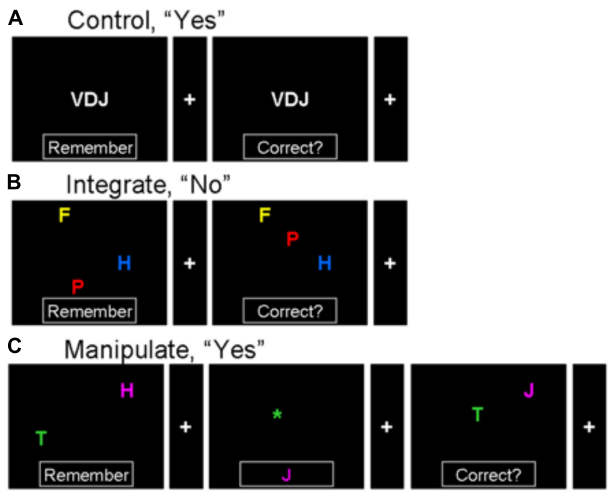


FIGURE 1 | Illustration of three kinds of cognitive tasks. Stimuli in the initial sample screen are encoded, and compared to a final probe, with an intervening "change cue" in the manipulate trials. **(A)** Sample then probe in control condition (with only letter identity maintenance); correct answer is "yes" since the letters in the sample are the same as the letters in the probe. **(B)** Integrate condition with maintenance of integrated letter identity, spatial and color information; a "no" trial since the identity, position, and color are not the same for each of the three letters in both sample and probe. **(C)** Sample, change cue, and probe in a manipulate condition trial requiring manipulation of integrated letter identity, spatial and color information; a "yes" trial. The two letters in the sample are encoded, and at the change cue the internal representation is altered so that the letter in the same color as the asterisk (green in this trial) is moved to the position of the asterisk, and the identity of the letter whose color matches the letter in the box is changed to match the letter in the box. This changed mental representation is compared to the final probe.

and tilted 45° to allow stimulus viewing. Response times and accuracy for participant responses to probe screens on each trial were recorded by the stimulus program.

Three different tasks or conditions were employed (see **Figure 1**). Each trial consisted of two or three sequentially presented screens: a sample (sample phase), blank or change cue (cue phase), and probe (retrieval phase). The first condition required maintenance of unintegrated information (control). Participants saw a three-letter sample stimulus, in which letters were all white and located in the center of a screen, with the word "Remember" in a box at the bottom of the screen. This was followed by an average 2-second-long inter-stimulus-interval (ISI) with a blank screen and then a fixation screen containing only a fixation cross and a blank box at the bottom. Next another roughly 2-s blank ISI was followed finally by a probe screen, again with three letters arranged in the middle. Participants were trained to indicate by pushing one of two buttons whether the letters in the sample were the same as the letters in this probe (the order of the letters did not matter, but in all "yes" trials the orders matched). Presentation of the sample, intervening fixation, and probe, with the participant's response, constituted a trial. In a second condition (integrate), three letters were in the sample, which were placed in random locations around the screen and presented using different colors randomly selected from red, blue, green, yellow, cyan and magenta. Again following a blank ISI a second screen contained a fixation and blank box at the bottom and another blank ISI, a probe containing three colored letters in different positions was presented. Participants indicated whether the letters in the probe matched those in the sample on letter identity, color, and position, requiring these features of each of the sample stimulus letters to be retained in integrated representations. In the third condition (manipulate), two colored and randomly positioned letters were in the sample and probe just as in the integrate condition but with one less letter, and the intervening screen contained one of two "change" cues along with the fixation cross. One of the change cues, an asterisk located somewhere on the screen, indicated that the letter matching the asterisk in color should be relocated to the position of the asterisk. The second change cue, a letter in the box at the bottom of the screen, indicated that the sample letter matching its color should be changed to that letter (see **Figure 1**). Thus, the "change" cue screen required subjects to change the identity of one of the two sample letters, and to change the location of the other. These manipulations were performed on the internally maintained representation of the sample stimuli. When the probe screen was presented, participants indicated whether the probe matched the new representation of the stimulus after manipulation in accordance with the change cue.

Following each trial, a 3–5 s ISI screen preceded the next trial. One-third of the time, a null event (2 s) and another ISI also intervened before the next trial. Stimulus duration for all sample, change, probe, and fixation screens was 2 s. If the participant did not respond to the probe within 5 s it was coded as an incorrect trial. ISI blank screen durations randomly varied from 3–5 s to jitter stimulus onsets throughout the experiment. Additionally, null events with 5–7 s' duration were presented between randomly selected trials. The study consisted of three runs, each 576 s long. During each run, 36 trials within each condition and 36 null events occurred in semi-random order.

Imaging Acquisition

Functional mages were acquired on a 3-Tesla Siemens Trio scanner located at the Mind Research Network in Albuquerque, New Mexico. T2*-weighted gradient echo, echo-planar images (EPI) comprised of 33 interleaved 3 mm-skip-1 mm slices parallel to the AC-PC line were acquired (TR = 2,000 ms, TE = 29 ms, Flip = 75°, FOV = 240 mm, Matrix = 64 × 64). Dummy volumes for 16 s initiated each run to equilibrate the signal and were discarded. A high-resolution T1 MPRAGE anatomical scan was also acquired.

Image Analyses

fMRI data analysis was performed using SPM5 (Wellcome Department of Cognitive Neurology, London, United Kingdom). Images were corrected for differences in slice timing by resampling all slices to match the middle slice using sinc interpolation (Henson et al., 1999). Corrected images then were spatially realigned to the first volume to correct head motion in each run of all subjects. No participant had moved more than 3 mm in any axis. The images were coregistered

with the anatomical image (MPRAGE) of each subject and then normalized to the standard T1 template (average 305) from the Montreal Neurological Institute (MNI). The images were resampled into 3 mm by normalization and spatially smoothed with an 8 mm FWHM isotropic Gaussian kernel. Data were high-pass filtered to remove low frequency noise with a 128 s cutoff period.

Statistical analyses were modeled using a canonical hemodynamic response function (HRF) and its derivatives. At the first level individual analysis, each event was calculated using an event-related design with all events including samples and cues of each task and null events. All task events were subtracted by null events and these contrast maps were used to analyze group data.

BOLD responses were compared between the samples for the control, integrate, and manipulate conditions. We also directly compared responses to the change cue of the manipulate condition (manipulate two integrated letters) with activations during the sample of the integrate condition (maintain three integrated letters) in order to reveal the differences in activation for manipulation and maintenance of integrated information. *P*-values then were cluster level corrected at $p < 0.05$. Based on group analyses, ROIs (10 mm spheres) were selected for further analyses and BOLD signal changes were extracted.

RESULTS

Behavioral Results

Mean accuracy and RT are depicted in **Figure 2**. We used one-way within-subjects ANOVAs to analyze accuracy and RT for task conditions. The accuracy was lower in the manipulate condition (0.726) than in control condition (0.926), $F(1, 14) = 46.460$, $p < 0.001$, and lower in the integrate condition (0.777) than in control condition, $F(1, 14) = 44.903$, $p < 0.001$. RT in the control condition (998 ms) was faster than in the integrate condition (1,222 ms), $F(1, 14) = 17.568$, $p < 0.001$, and faster than in the manipulate condition (1,240 ms), $F(1, 14) = 39.482$, $p < 0.001$.

fMRI Results

Sample Phase

We first contrasted activation for the sample phases for the control, integrate, and manipulate conditions. **Supplementary Table 1** lists coordinates and activations for local maxima for which there were significant differences in BOLD responses for all contrasts performed. Activations for these contrasts are illustrated in **Figure 3** along with time courses of the activations. Contrasting neural responses to the sample in the integrate condition, when subjects encoded three colored letters in random locations, to the sample in the control condition, when subjects encoded three centered, white letters, a broad network of regions were more activated by the sample of the integrate condition. This included left and right inferior and middle frontal gyrus (BA 6, 9, and left 46), left superior frontal gyrus (BA 6), left and right precentral gyrus (BA 6), and left insula (BA 13). Medially, anterior cingulate (BA 32), cingulate gyrus (BA 24), and right cuneus (BA 17) were more strongly activated for the integrate sample. Posteriorly, right superior parietal (BA 7), bilateral inferior parietal lobule (BA 40), and bilateral precuneus were activated, along with right superior temporal gyrus (BA 22), left and right middle occipital gyrus (BA 19), bilateral lingual gyrus (BA 17/18), and left fusiform gyrus (BA 37). Subcortically, bilateral caudate, right claustrum, and lentiform nucleus were also recruited more in the integrate sample than the control sample, as were right thalamus and bilateral cerebellum. A similar network was more active during the sample phase of the manipulate condition, when subjects viewed two letters of different colors and in random locations, with the exceptions of the right inferior gyrus, right precentral gyrus, and BA 24 in cingulate gyrus.

Notably, the ROI in left DLPFC (BA 9) was larger in volume in the manipulate and integrate samples than the maintenance sample period, extending ventrally to Talairach coordinates −47, −6, 34 during manipulation and −47, 3, 11 during the integrate sample. Two regions in DLPFC that were significant in the manipulate sample were absent in the integrate sample contrast, despite the greater amount of integrated information in the integrate sample. These ROIs were fairly anterior in BA 9

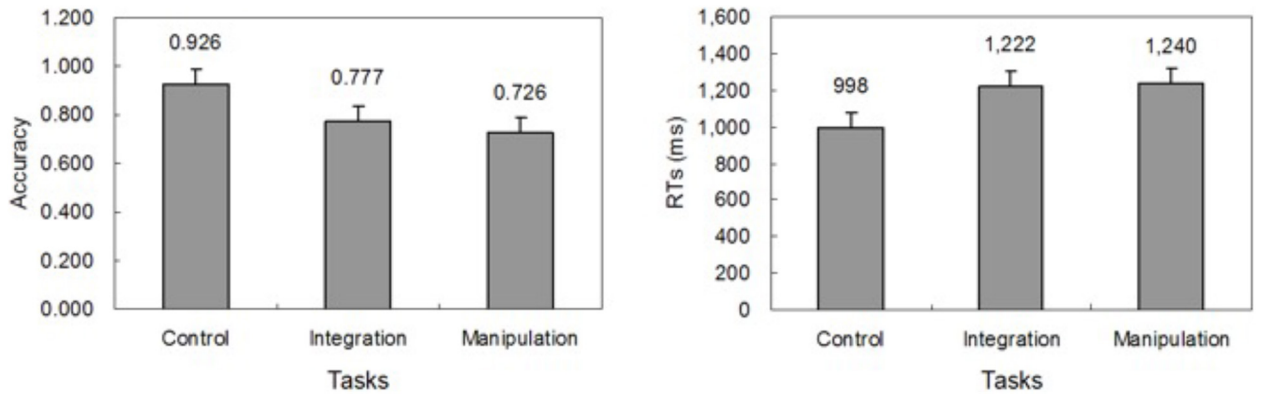


FIGURE 2 | Behavioral results. Mean accuracy (left) and response time (right) across subjects for the control, integrate, and manipulate conditions.

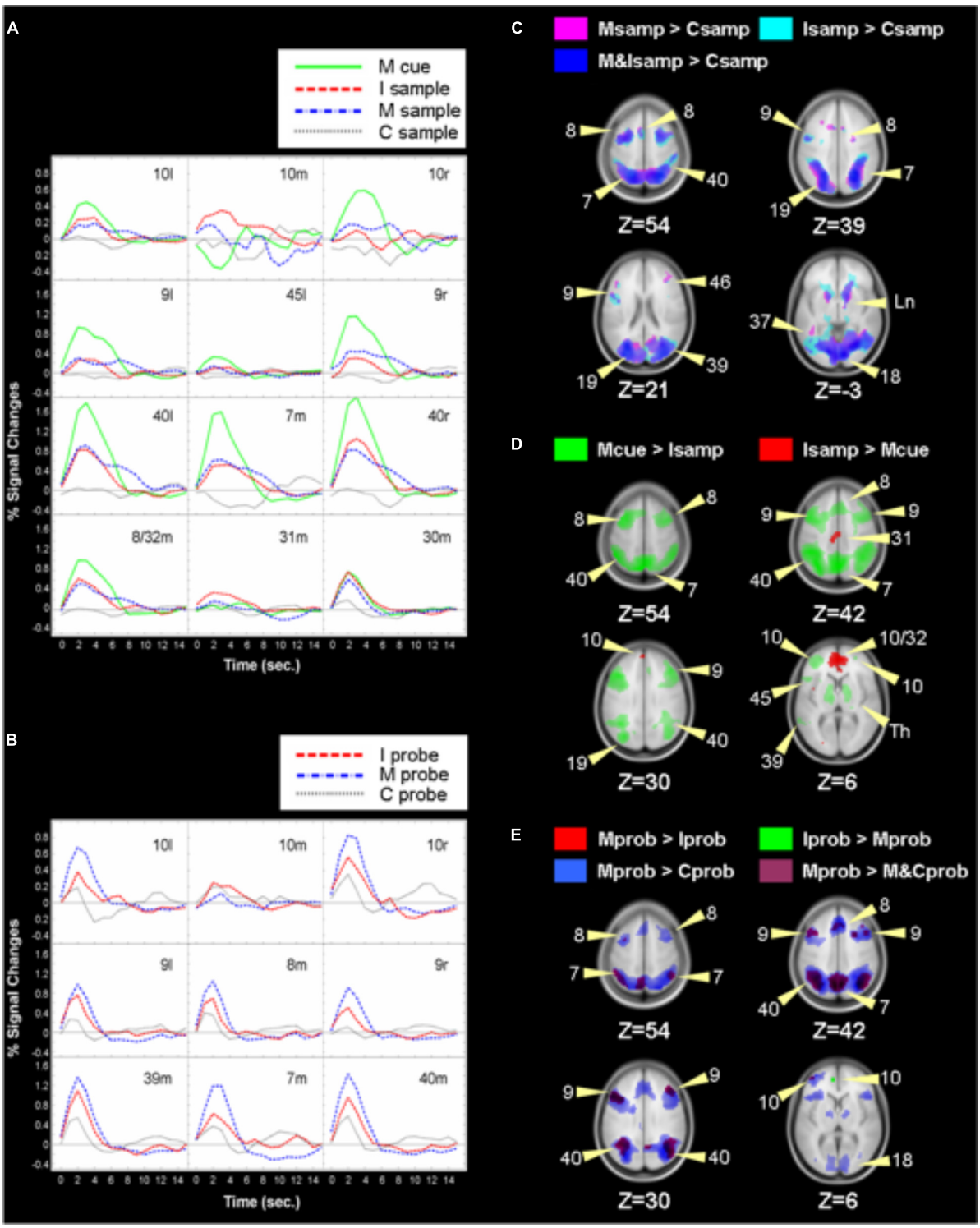


FIGURE 3 | Labels refer to Brodmann's areas and left or right hemisphere or medial. For example, in panel **(A)**, 10l refers to left Brodmann Area 10 in the left hemisphere. The green line represents activation or BOLD intensity in that area during the manipulate trials' cue phase. We compared responses for trial phases and depicted them in panels **(C–E)**, with labeled arrows indicating Brodmann Areas. The subtraction of the integrate trial sample phase from the manipulate trial change cue phase resulted in significant activations shown in green in panel **(D)**. The opposite subtraction produced the medial red activations. The green ROIs' activations correspond to the green lines in panel **(A)**. Thus, activations in panels **(A,B)** can be seen to arise from the same-colored ROIs in panels **(C–E)**. Depicted in panel **(A)**: BOLD responses during sample and change cue phases of control, integrate, and manipulate trials for ROIs depicted in panels **(C,D)**. **(B)** BOLD responses for ROIs resulting from contrasting the probe phase in control, integrate, and manipulate trials, as depicted in panel **(E)**. C, Control, I, Integrate, M, Manipulate trial types. Samp, response to the sample phase of a trial type, cue, response to the change cue, prob, response to the probe phase.

(Talairach coordinates −46, 18, 9 and 38, 38, 33). We speculate that this may have reflected activity as participants prepared for or anticipated the impending change cue in the manipulate trials. Given that subjects were to manipulate the stimuli in the manipulate sample, these frontal activations may reflect strategy or preparation processing. There were no significant differences between activation evoked by the manipulate and integrate samples, nor were any regions more active during the control sample than during the integrate or manipulate samples.

Probe Phase

Second, we contrasted responses to the probe stimuli in the control, integrate, and manipulate conditions. A similar network of regions was more significantly active during the integrate probe and the manipulate probe than during the control probe. However, in parietal cortex (BA 39/40), a larger volume ROI was evoked by the manipulate probe than the integrate probe, and a larger amplitude response was evoked in middle occipital cortex during the integrate probe. Parietal cortex is active for visual imagery, particularly manipulation of visual imagery (Kosslyn et al., 2001); this activation during the probe suggests these circuits were active to maintain the manipulated stimulus while compared to the probe. Response time was slower for the manipulate probe even though two maintained letters were compared to two probe letters in that condition compared to three in the integrate condition. Comparing the participant-created representation was more demanding than comparing the larger encoded representation to the probes. Greater activation in mid-occipital regions probably reflects the greater demand on visual working memory to maintain the larger stimulus. These results together suggest that while the integrate sample was retained in visual working memory, representing the stimulus generated by participants depended less on visual substrates.

The manipulate condition probe was contrasted directly with the integrate condition probe, revealing a cortical network more active in the manipulate probe including left frontal pole (BA10), bilateral DLPFC (left and right BA 9, left BA6), medial frontal gyrus (BA 8/32), left precuneus (BA7), left angular gyrus and right middle temporal gyrus (BA 39), left claustrum, and bilateral pyramids of the cerebellum. The reverse contrast revealed a significant difference only in the anterior cingulate (left BA32). Though the integrate condition probe involved operations with more extensive integrated representations, performing the same operations on generated stimuli involved greater activity in frontal cortex, particularly including frontal pole.

Manipulate Cue and Samples

Next we compared neural responses to manipulation of two stimulus letters and to encoding of three stimulus letters in the control and integrate condition samples. Activity was significantly greater during manipulation than for the control sample across cortical regions including BA 6, 9, and 46 in frontal cortex and 7, 40, and 9 posteriorly. Activation in response to the manipulation change cue was then contrasted with activity for the integrate sample. This comparison addresses the primary aim of this study: contrasting simple representation of integrated information, and manipulation of it. These stimuli differed in that the manipulation cue entailed both integration of features and manipulation of the integrated representation of two colored and randomly positioned letters while the integrate sample entailed encoding and retention of three of them. Comparing activity in response to the manipulation cue and the integrate sample allowed us to isolate activation specific to manipulation of an integrated representation as both required representation of integrated information. In fact, since the integrate sample contained three items and the manipulation stimulus contained two, the demand on working memory capacity to sustain the representation of the integrate sample was greater than for the manipulation cue, yet a network of regions similar to the manipulation cue minus the control sample responded more to the manipulation cue than the integrate sample. An exception is that activation in occipital visual areas was evident when contrasting the manipulate cue to the control sample, but not in the contrast between the manipulate cue and the integrate sample, suggesting that maintaining the integrated representation of the three letters in the integrate sample and manipulating the integrated representation of two letters depended on the same visual processing regions.

The time courses of BOLD responses were plotted for ROIs that activated significantly more to the manipulate cue than to the integrate sample phase. For each of these ROIs, a plot in **Figure 3** depicts the time courses of BOLD responses for the sample phases of each condition, and the manipulate condition change cue. Lateral FPC, especially in the right hemisphere, responded more to manipulation than during the samples. Though not significant, the time courses suggest some participation of FPC during the integrate and manipulate samples mostly in left FPC. Responses to the manipulate cue were sharply increased in DLPFC and parietal cortex relative to all of the sample periods. This network accomplishing the manipulation exhibited strong dependence on FPC while sustaining representations of the sample stimuli did not. A graded increase in response intensity from the control sample to the integrate sample and to the manipulate cue is seen in several of the ROIs across cortex. Responses to the control sample were surprisingly small, since activations for DLPFC are typically found for match-to-sample paradigms. The contrasts applied may have failed to produce ROIs where activation during the control condition sample occurred. These regions did, however, respond during the control probe.

Left inferior frontal gyrus (BA 45, Broca's area) activated for integration and more for manipulation but was slightly suppressed during the control sample. Bilateral frontal eye fields (BA 6) and supplementary eye field (BA 8/32) showed graded responses to the integrate sample and manipulation but were not responsive to the control sample. The most intense responses to integration and especially manipulation were observed in parietal cortex, in medial superior precuneus (BA7) and bilateral inferior parietal cortex (BA 39/40, but bordering in lateral BA7).

Analysis also revealed four maxima that were more active during the integrate sample than during the manipulation. These areas were significantly different due to combinations of activation to integrating and suppression during manipulation. They included medial FPC (BA 10), anterior cingulate (BA 32), and both dorsal (BA 31) and posterior cingulate (BA 30).

Lateral frontal pole, especially in the right hemisphere, responded strongly to manipulation, but a decrease below baseline occurred in medial frontal pole during manipulation. Different patterns of lateral and medial responses occurred during the sample for the control and integrate conditions. Neither medial nor lateral frontal pole appears to have participated in encoding of the control sample. Both left and medial FPC were recruited during the integrate sample, but little activity was evident in right FPC. These BOLD plots suggest very different engagement of medial and lateral frontal pole during integration and manipulation; the relationship is enhanced by deactivation in medial FPC for manipulation. To gauge the active contribution of these areas during the tasks, we obtained from each subject the maximum BOLD activation for these regions following stimulus presentation for the integrate sample and manipulate change cue phases. From these the average peak activation across subjects for the two kinds of trial phase were determined and are plotted in **Figure 4**; these reflect the greatest activation of these regions and are not influenced by deactivations. There is an interaction [$F(1, 14) = 6.800$, $p < 0.05$] between points for the medial and right FPC maxima, but not when all points are considered. This result makes clear that even discounting deactivations in BOLD, response integration and manipulation produce a different pattern of recruitment across lateral and medial FPC.

To assess the relationship between observed BOLD differences and performance on the experimental paradigm, correlations between activation level for each ROI and mean response time were computed. For the integrate trials, response time correlated negatively with activation in left DLPFC during the sample (BA 9, $r = -0.52$) and with activation in left DLPFC and left inferior parietal cortex during the probe (BA 9 and 39, $r = -0.56$ and -0.61, **Figure 5**). Activation during the manipulate cue in a network including right FPC, left DLPFC, and right inferior parietal lobe appears able to account for accuracy in the manipulate trials, while increased left DLPFC and inferior parietal activity resulted in faster responses on integrate trials, possibly indicating additional effort in these regions during encoding and solution.

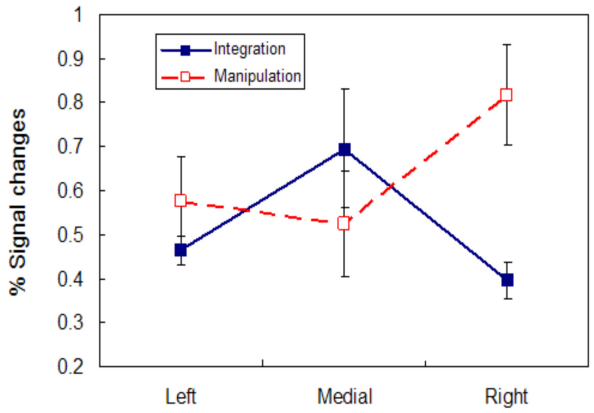


FIGURE 4 | Lateral and medial FPC activity during integration and manipulation. Mean peak activity represented as percent change of the BOLD signal in left, medial, and right FPC during the integrate sample (solid line) and manipulate condition change cue (dashed line). Mean maximum activation amplitude across subjects is plotted to assess processing in each region without influence from deactivations.

DISCUSSION

The purpose of this study was to separate and compare the demands placed on frontopolar cortex by representation of integrated information and manipulation of integrated information. The information integrated, letter identity, color, and location, formed through the integration complex, structured representations. Specifically, we manipulated the degree of integration complexity of representations in working memory, and the requirement to change the information to produce a novel integrated representation. Maintaining an integrated representation of three letters presented in three different colors and locations did not recruit FPC significantly more than the control condition requiring maintenance of three letters for which only the identity of the letters was pertinent. FPC was recruited, however, when a smaller structured representation integrating two letters in different colors and locations was manipulated. This region was significantly more active for manipulation of the two-letter stimulus than for maintenance of the integrated 3-letter sample, and than for maintenance of the simple sample in the control condition. These findings suggest that FPC activation found for paradigms involving integration of information results from the need to manipulate or create integrated information, rather than the demands of representing integrated information.

Other studies have shown that internal manipulation of information alone is not sufficient to recruit FPC, for example, in N-back paradigms (Cohen et al., 1997) or math performance (Dehaene et al., 1999). This indicates that manipulation alone is not a sufficient demand to recruit FPC. In our manipulate condition, letter identity and color must be bound together with a location. Each letter's position was defined by its spatial relationship to the other letters in the stimulus and the surrounding frame. These constraints structure the encoded representation of the stimulus. Maintaining this constrained representation in working memory was insufficient to recruit FPC; it was recruited in this study only when novel integrated information was produced.

A key element of our experimental design is that the cognitive demands of the integrated representation in the integrate condition (3 letters) was larger than that in the manipulate condition (2 letters). It is possible that during manipulation intermediate representations were employed in which stimulus features progressively changed, which when combined with retaining the sample stimuli until the manipulation was complete, summed to demand more integration than in our integrate condition. In this interpretation, FPC activation in response to the change cue may be a result of holding a sufficient amount of information in integrated form. We suggest that the production of these representations according to task constraints and the representations thus produced essentially are manipulation. It is this formative process that we propose is the fundamental contribution of FPC.

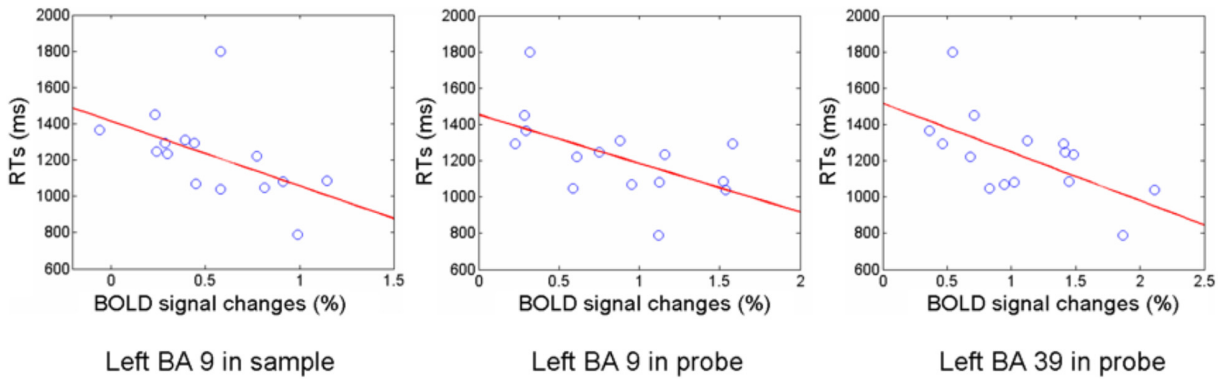


FIGURE 5 | RT and activation amplitude correlations. Correlation between reaction time and brain activation in BA9 in the sample phase of the integrate task, and in BA 9 and BA 39 during the probe phase of the integrate task.

It might also be argued that when the cue was presented for 2 s, containing a colored asterisk in some location and a letter in the box at the bottom of the stimulus screen, there was a demand for the participant to integrate the original two sample letters, and their locations and colors, as well as the asterisk's color and location, and the identity of the letter in the box with its color. Formally, the relational complexity of this representation is smaller than that demanded by the integrate sample which entailed three letter identities, three locations, and three colors [see Holyoak and Thagard (1995) and Halford et al. (2007) for discussion of the formalization of complexity degree]. For the manipulate cue, in addition to the two sample letters each requiring binding a letter identity, color, and location, a position was bound to a color (asterisk) and a letter identity was bound to a color (letter in the box), so less integration was required than for maintaining the integrate sample. Therefore the size of the representation explicitly required by the cue contained fewer bindings than the integrate sample. Pragmatically, the cue was present for 2 s, during which it is likely that at least part of the manipulation was completed, reducing the need to retain the change cues in working memory, further reducing the degree of integration required. The manipulation performed in fact results in constituting a new integration of elements of the sample and change cue. The distinction between this and the mental activity occurring during the integrate sample speaks to the essential aim of this study—that manipulation of integrated representations involves production of additional integrated information. The integrated nature of the information constrains constitution of new representations. We propose that this constrained production of representations is the ideal sort of cognition to be served by the integrative physical character of FPC. There is no obvious theoretical reason why this description of neural processing should be restricted to information about external stimuli, information about task execution, or about relative reward associated with action possibilities, all of which may be constrained to arbitrary levels of complexity. In this view, managing multiple distinct representations adds both information and complexity. Thus, it might be possible to observe greater FPC activity for a single, complexly constrained manipulation than multiple simpler ones, and simple manipulations upon a complexly constrained representation might produce similar demand to complex manipulations of relatively simple information. These theoretical proposals may be easily co-opted into testable hypotheses. The multiplicity of paradigms which produce FPC activation as a body witness the flexibility of constrained production.

Frontopolar cortex—the same FPC region more activated for manipulation than the integrate or control samples—was also recruited during the probe phases, as is depicted in **Figure 3**. Left FPC responded significantly more to the integrate probe and manipulate probes than the maintenance probe (not shown) and to the manipulate probe than the integrate probe. Right FPC also attained significance when the manipulate probe was compared to the control probe, and as seen in the BOLD time courses was more active than left FPC. Whereas neither the integrate nor manipulate sample periods recruited FPC relative to the control sample, the probe phase for both of those conditions recruited FPC more than in the control. As is also apparent in the time courses in **Figure 3**, comparing an integrated representation of two letters which had been produced by participants then retained for several seconds to the letters in the probe recruited FPC more than comparing three perceived and encoded integrated letters to three letters in the probe. The nominal cognitive load is greater in the latter case, but when the smaller representation had been created by the participant, the comparison depended much more on FPC—again, we propose, exploiting the integrative anatomical character and connectivity of FPC to sustain the participant-produced representation. This demand on FPC results from the need to maintain the produced representation without any memory of a perceived stimulus to refer to.

The recruitment of right FPC during the manipulate cue and probe may result simply because the task entailed greater integration demand and relied upon more of FPC, or because of functional specialization in right FPC. Spatial processing has been associated with the right hemisphere (Kosslyn et al., 1994; Baddeley, 1996; Smith et al., 1996; Manoach et al., 2004). Slotnick and Moo (2006) showed that memory for coordinate

location (a dot was far from a figure) recruited right FPC, while categorical memory (the dot is on the figure) recruited left FPC. Manipulating letters at the change cue entailed manipulating position in coordinate space.

Humans operate within complex environments comprised of complex information. To select action in service of their goals in novel situations requires the ability to create plans from existing information. The central question of this study is whether FPC augments human cognitive ability by enabling representation of complex information, or whether it facilitates processing of complex information into new structured representations. The results support the latter conclusion. Even though a greater quantity of information had to be integrated in the integrate trials than in the manipulate trials, FPC was recruited only when changes were made to the representation to create a new representation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of New Mexico IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Both authors contributed equally and approved the article for publication.

FUNDING

This work was supported by NIH grant number S06 GM008136.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnsys.2022.788395/full#supplementary-material

REFERENCES

Andrews, G., and Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognit. Psychol.* 45, 153–219. doi: 10.1016/s0010-0285(02)00002-6

Baddeley, A. (1996). The fractionation of working memory. *Proc. Natl. Acad. Sci. U S A* 93, 13468–13472.

Badre, D., and Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends Cogn. Sci.* 22, 170–188. doi: 10.1016/j.tics.2017.11.005

Badre, D., and Wagner, A. D. (2004). Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron* 41, 473–487. doi: 10.1016/s0896-6273(03)00851-1

Barde, L. H., and Thompson-Schill, S. L. (2002). Models of functional organization of the lateral prefrontal cortex in verbal working memory: evidence in favor of the process model. *J. Cogn. Neurosci.* 14, 1054–1063. doi: 10.1162/089892902320474508

Bazargani, N. Y., Hillebrandt, H., Christoff, K., and Dumontheil, I. (2014). Developmental changes in effective connectivity associated with relational reasoning. *Hum. Brain Map.* 2014:35. doi: 10.1002/hbm.22400

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., and Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. *Neuron* 33, 301–311. doi: 10.1016/s0896-6273(01)00583-9

Bunge, S. A., Helskog, E. H., and Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage* 46, 338–342. doi: 10.1016/j.neuroimage.2009.01.064

Christoff, K., and Gabrieli, J. D. (2000). The frontopolar cortex and human cognition: Evidence for rostrocaudal hierarchial organization within the human prefrontal cortex. *Psychobiology* 28, 168–186. doi: 10.1093/cercor/bhu311

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., et al. (2001). Rostrolateral Prefrontal Cortex Involvement in Relational Integration during Reasoning. *NeuroImage* 14, 1136–1149. doi: 10.1006/nimg.2001.0922

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., et al. (1997). Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604–608. doi: 10.1038/386604a0

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114. doi: 10.1017/s0140525x01003922

Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., and Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Dev. Sci.* 12, 55–66. doi: 10.1111/j.1467-7687.2008.00743.x

De Pisapia, N., Slomski, J. A., and Braver, T. S. (2007). Functional specializations in lateral prefrontal cortex associated with the integration and segregation of information in working memory. *Cereb Cortex* 17, 993–1006.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., and Tsivkin, S. (1999). Sources of mathematical thinking: behavioral and brain-imaging evidence. *Science* 284, 970–974. doi: 10.1126/science.284.5416.970

D'Esposito, M., Postle, B. R., Ballard, D., and Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. *Brain Cogn.* 41, 66–86. doi: 10.1006/brcg.1999.1096

Dixon, M. L., Girn, M., and Christoff, K. (2017). "Hierarchical Organization of Frontoparietal Control Networks Underlying Goal-Directed Behavior," in *The Prefrontal Cortex as an Executive, Emotional, and Social Brain*, ed. M. Watanabe (New York, NY: Springer), 133–148. doi: 10.1007/978-4-431-56508-6_7

Duverne, S., and Koechlin, E. (2017). *Hierarchical Control of Behaviour in Human Prefrontal Cortex*. Hoboken, N J: Wiley.

Eichenbaum, A., Scimeca, J. M., and Esposito, D. (2020). Dissociable Neural Systems Support the Learning and Transfer of Hierarchical Control Structure. *J. Neurosci.* 40, 6624–6637. doi: 10.1523/JNEUROSCI.0847-20.2020

Fangmeier, T., Knauff, M., Ruff, C. C., and Sloutsky, V. (2006). FMRI evidence for a three-stage model of deductive reasoning. *J. Cogn. Neurosci.* 18, 320–334. doi: 10.1162/089892906775990651

Flechsig, P. (1901). Developmental (myelogenetic) localisation of the cerebral cortex in the human subject. *Lancet* 2, 1027–1029.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci. U S A* 101, 8174–8179. doi: 10.1073/pnas.0402680101

Goldberg, R. F., Perfetti, C. A., Fiez, J. A., and Schneider, W. (2007). Selective retrieval of abstract semantic knowledge in left prefrontal cortex. *J. Neurosci.* 27, 3790–3798. doi: 10.1523/JNEUROSCI.2381-06.2007

Grasby, P. M., Frith, C. D., Friston, K. J., Simpson, J., Fletcher, P. C., Frackowiak, R. S., et al. (1994). A graded task approach to the functional mapping of brain areas implicated in auditory-verbal memory. *Brain* 117(Pt 6), 1271–1282. doi: 10.1093/brain/117.6.1271

Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., and Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Res.* 1096, 125–137. doi: 10.1016/j.brainres.2006.04.024

Halford, G. S., Cowan, N., and Andrews, G. (2007). Separating cognitive capacity from knowledge: a new hypothesis. *Trends Cogn. Sci.* 11, 236–242. doi: 10.1016/j.tics.2007.04.001

Henson, R. N., Buchel, C., Josephs, O., and Friston, K. J. (1999). The slice-timing problem in event-related fMRI. *Neuroimage* 9:125.

Holyoak, K. J., and Thagard, P. (1995). *Mental Leaps: Analogy in Creative Thought*. Bradford, PA: Bradford Books.

Jacobs, B., Driscoll, L., and Schall, M. (1997). Life-span dendritic and spine changes in areas 10 and 18 of human cortex: a quantitative Golgi study. *J. Comp. Neurol.* 386, 661–680. [pii] doi: 10.1002/(SICI)1096-9861(19971006)386:4<661::AID-CNE11<3.0.CO;2-N

Jacobs, B., Schall, M., Prather, M., Kapler, E., Driscoll, L., Baca, S., et al. (2001). Regional dendritic and spine variation in human cerebral cortex: a quantitative golgi study. *Cereb Cortex* 11, 558–571. doi: 10.1093/cercor/11.6.558

Jerath, R., Beveridge, C., and Jensen, M. (2019). On the Hierarchical Organization of Oscillatory Assemblies: Layered Superimposition and a Global Bioelectric Framework. *Front. Hum. Neurosci.* 2019:13. doi: 10.3389/fnhum.2019.00426

Johnson, M. K., Raye, C. L., Mitchell, K. J., Greene, E. J., Cunningham, W. A., and Sanislow, C. A. (2005). Using fMRI to investigate a component process of reflection: prefrontal correlates of refreshing a just-activated representation. *Cogn. Affect Behav. Neurosci.* 5, 339–361. doi: 10.3758/cabn.5.3.339

Korb, F. M., Jiang, J., King, J. A., and Egner, T. (2017). Hierarchically Organized Medial Frontal Cortex-Basal Ganglia Loops Selectively Control Task- and Response-Selection. *J. Neurosci.* 37, 7893–7905. doi: 10.1523/JNEUROSCI.3289-16.2017

Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Chabris, C. F., Rauch, S. L., and Anderson, A. K. (1994). Identifying objects seen from different viewpoints. A PET investigation. *Brain* 117(Pt 5), 1055–1071. doi: 10.1093/brain/117.5.1055

Kosslyn, S. M., Ganis, G., and Thompson, W. L. (2001). Neural foundations of imagery. *Nat. Rev. Neurosci.* 2, 635–642. doi: 10.1038/35090055

Krawczyk, D. C., McClelland, M. M., and Donovan, C. M. (2011). A hierarchy for relational reasoning in the prefrontal cortex. *Cortex* 47, 588–597. doi: 10.1016/j.cortex.2010.04.008

Kroger, J. K., Holyoak, K. J., and Hummel, J. E. (2004). Varieties of sameness: the impact of relational complexity on perceptual comparisons. *Cogn. Sci.* 28, 335–358.

Kroger, J. K., Nystrom, L., Cohen, J. D., and Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Res.* 1243, 86–103. doi: 10.1016/j.brainres.2008.07.128

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cort.* 12, 477–485. doi: 10.1093/cercor/12.5.477

Loewenstein, J., and Gentner, D. (2005). Relational language and the development of relational mapping. *Cogn. Psychol.* 50, 315–353. doi: 10.1016/j.cogpsych.2004.09.004

MacLeod, A. K., Buckner, R. L., Miezin, F. M., Petersen, S. E., and Raichle, M. E. (1998). Right anterior prefrontal cortex activation during semantic monitoring and working memory. *Neuroimage* 7, 41–48. doi: 10.1006/nimg.1997.0308

Manoach, D. S., White, N. S., Lindgren, K. A., Heckers, S., Coleman, M. J., Dubal, S., et al. (2004). Hemispheric specialization of the lateral prefrontal cortex for strategic processing during spatial and shape working memory. *Neuroimage* 21, 894–903. doi: 10.1016/j.neuroimage.2003.10.025

Mansouri, F. A., Freedman, D. J., and Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* 21, 595–610. doi: 10.1038/s41583-020-0364-5

Mestres-MissÈ, A., Turner, R., and Friederici, A. D. (2012). An anterior posterior gradient of cognitive control within the dorsomedial striatum. *NeuroImage* 62, 41–47. doi: 10.1016/j.neuroimage.2012.05.021

Monti, M. M., Osherson, D. N., Martinez, M. J., and Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage* 37, 1005–1016. doi: 10.1016/j.neuroimage.2007.04.069

Nee, D. E., and D Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *eLife* 2016:5. doi: 10.7554/eLife.\break12112

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* 331, 585–589. doi: 10.1038/331585a0

Petrides, M., Alivisatos, B., Meyer, E., and Evans, A. C. (1993). Functional activation of the human frontal cortex during the performance of verbal working memory tasks. *Proc. Natl. Acad. Sci. U S A* 90, 878–882. doi: 10.1073/pnas.90.3.878

Pollmann, S., Weidner, R., Mller, H. J., and Cramon, D. Y. V. (2000). A Fronto-Posterior Network Involved in Visual Dimension Changes. *J. Cogn. Neurosci.* 12, 480–494. doi: 10.1162/089892900562156

Prabhakaran, V., Narayanan, K., Zhao, Z., and Gabrieli, J. D. (2000). Integration of diverse information in working memory within the frontal lobe. *Nat. Neurosci.* 3, 85–90. doi: 10.1038/71156

Reynolds, J. R., McDermott, K. B., and Braver, T. S. (2006). A direct comparison of anterior prefrontal cortex involvement in episodic retrieval and integration. *Cereb Cortex* 16, 519–528. doi: 10.1093/cercor/bhi131

Riddle, J., Vogelsang, D. A., Hwang, K., Cellier, D., and D Esposito, M. (2020). Distinct Oscillatory Dynamics Underlie Different Components of Hierarchical Cognitive Control. *J. Neurosci.* 40, 4945–4953. doi: 10.1523/JNEUROSCI.0617-20.2020

Rusu, S. I., and Pennartz, C. M. A. (2019). Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized cortico basal ganglia systems. *Hippocampus* 30, 73–98. doi: 10.1002/hipo.23167

Rypma, B., Prabhakaran, V., Desmond, J. E., Glover, G. H., and Gabrieli, J. D. (1999). Load-dependent roles of frontal brain regions in the maintenance of working memory. *Neuroimage* 9, 216–226. doi: 10.1006/nimg.1998.0404

Sarafyazd, M., and Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. *Science* 2019:364. doi: 10.1126/science.aav8911

Schumacher, F. K., Schumacher, L. V., Schelter, B., and Kaller, C. P. (2019). Functionally dissociating ventro-dorsal components within the rostro-caudal hierarchical organization of the human prefrontal cortex. *NeuroImage* 185, 398–407. doi: 10.1016/j.neuroimage.2018.10.048

Segalowitz, S. J., and Davies, P. L. (2004). Charting the maturation of the frontal lobe: an electrophysiological strategy. *Brain Cogn.* 55, 116–133. doi: 10.1016/S0278-2626(03)00283-5

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., and Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: a comparative study of area 10. *Am. J. Phys. Anthropol.* 114, 224–241. [pii] doi: 10.1002/1096-8644(200103)114:3<224::AID-AJPA1022<3.0.CO;2-I

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature* 440, 676–679. doi: 10.1038/nature04513

Slotnick, S. D., and Moo, L. R. (2006). Prefrontal cortex hemispheric specialization for categorical and coordinate visual spatial memory. *Neuropsychologia* 44, 1560–1568. doi: 10.1016/j.neuropsychologia.2006.01.018

Smith, E. E., Jonides, J., and Koeppe, R. A. (1996). Dissociating verbal and spatial working memory using PET. *Cereb. Cortex* 6, 11–20. doi: 10.1093/cercor/6.1.11

Strange, B. A., Henson, R. N., Friston, K. J., and Dolan, R. J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cereb. Cortex* 11, 1040–1046. doi: 10.1093/cercor/11.11.1040

Sweeney, J. A., Mintun, M. A., Kwee, S. B., Wiseman, M., Brown, D. L., Rosenberg, D. R., et al. (1996). Positron emission tomography study of voluntary saccadic

eye movements and spatial working memory. *J. Neurophys.* 75, 454–468. doi: 10.1152/jn.1996.75.1.454

Thiebaut de Schotten, M., Urbanski, M., Batrancourt, B., Levy, R., Dubois, B., et al. (2017). Rostro-caudal Architecture of the Frontal Lobes in Humans. *Cereb. Cortex* 27:4047. doi: 10.1093/cercor/bhw215

Uttal, D. H., Gentner, D., Liu, L. L., and Lewis, A. R. (2008). Developmental changes in children's understanding of the similarity between photographs and their referents. *Dev. Sci.* 11, 156–170. doi: 10.1111/j.1467-7687.2007.00660.x

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., et al. (2015). Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nat. Neurosci.* 18, 1318–1324. doi: 10.1038/nn.4071

Wendelken, C., Nakhabenko, D., Donohue, S. E., Carter, C. S., and Bunge, S. A. (2008). Brain Is to Thought as Stomach Is to ??: Investigating the Role of Rostrolateral Prefrontal Cortex in Relational Reasoning. *J. Cogn. Neurosci.* 20, 682–693. doi: 10.1162/jocn.2008.20055

Wood, J. N., and Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nat. Rev. Neurosci.* 4, 139–147. doi: 10.1038/nrn1033

Yufik, Y. (2019). The Understanding Capacity and Information Dynamics in the Human Brain. *Entropy* 2019:21. doi: 10.3390/e21030308

Yufik, Y. M., and Friston, K. J. (2016). Life and Understanding: the Origins of@ Understanding in Self-Organizing Nervous Systems. *Front. Syst. Neurosci.* 2016:10. doi: 10.3389/fnsys.2016.00098

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 10 March 2022
doi: 10.3389/fnsys.2022.764708

Integrating Philosophy of Understanding With the Cognitive Sciences

Kareem Khalifa[1], Farhan Islam[2], J. P. Gamboa[3], Daniel A. Wilkenfeld[4] and Daniel Kostić[5]*

[1] Department of Philosophy, Middlebury College, Middlebury, VT, United States, [2] Independent Researcher, Madison, WI, United States, [3] Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, United States, [4] Department of Acute and Tertiary Care, University of Pittsburgh School of Nursing, Pittsburgh, PA, United States, [5] Institute for Science in Society (ISiS), Radboud University, Nijmegen, Netherlands

We provide two programmatic frameworks for integrating philosophical research on understanding with complementary work in computer science, psychology, and neuroscience. First, philosophical theories of understanding have consequences about how agents should reason if they are to understand that can then be evaluated empirically by their concordance with findings in scientific studies of reasoning. Second, these studies use a multitude of explanations, and a philosophical theory of understanding is well suited to integrating these explanations in illuminating ways.

Keywords: explanation, understanding, mechanism, computation, topology, dynamic systems, integration

OPEN ACCESS

Edited by:
Yan Mark Yufik, Virtual Structures Research Inc., United States

Reviewed by:
Raoul Gervais, University of Antwerp, Belgium
Marcin Miłkowski, Institute of Philosophy and Sociology (PAN), Poland

****Correspondence:***
Kareem Khalifa
kkhalifa@middlebury.edu

Received: *25 August 2021*
Accepted: *10 February 2022*
Published: *10 March 2022*

Citation:
Khalifa K, Islam F, Gamboa JP, Wilkenfeld DA and Kostić D (2022) Integrating Philosophy of Understanding With the Cognitive Sciences. Front. Syst. Neurosci. 16:764708. doi: 10.3389/fnsys.2022.764708

INTRODUCTION

Historically, before a discipline is recognized as a science, it is a branch of philosophy. Physicists and chemists began their careers as "natural philosophers" during the Scientific Revolution. Biology and psychology underwent similar transformations throughout the nineteenth and early twentieth centuries. So, one might think philosophical discussions of understanding will be superseded by a "science of understanding."

While we are no great forecasters of the future, we will suggest that philosophical accounts of understanding can make two important scientific contributions. First, they provide a useful repository of hypotheses that can be operationalized and tested by scientists. Second, philosophical accounts of understanding can provide templates for unifying a variety of scientific explanations.

We proceed as follows. We first present these two frameworks for integrating philosophical ideas about understanding with scientific research. Then we discuss the first of these frameworks, in which philosophical theories of understanding propose hypotheses that are tested and refined by the cognitive sciences. Finally, we discuss the second framework, in which considerations of understanding provide criteria for integrating different scientific explanations. Both of our proposals are intended to be programmatic. We hope that many of the relevant details will be developed in future work.

TWO FRAMEWORKS FOR INTEGRATION

As several reviews attest (Baumberger, 2014; Baumberger et al., 2016; Gordon, 2017; Grimm, 2021; Hannon, 2021), understanding has become a lively topic of philosophical research over

the past two decades. While some work has been done to integrate these ideas with relevant findings from computer science, psychology, and neuroscience, these interdisciplinary pursuits are relatively nascent. While other frameworks are possible and should be developed, we propose two ways of effecting a more thoroughgoing synthesis between philosophy and these sciences (**Figure 1**). In the first framework for integrating philosophy with the cognitive sciences—what we call *naturalized epistemology of understanding* (**Figure 1A**)—the philosophy of understanding provides conjectures about reasoning that are tested and explained by the relevant sciences. In the second integrative framework—*understanding-based integration* (**Figure 1B**)—the philosophy of understanding provides broad methodological guidelines about how different kinds of scientific explanation complement each other. The two proposals are independent of each other: those unpersuaded by one may still pursue the other. We discuss each in turn.

NATURALIZED EPISTEMOLOGY OF UNDERSTANDING

In epistemology, naturalism is the position that philosophical analyses of knowledge, justification, and kindred concepts should be intimately connected with empirical science. Different naturalists specify this connection in different ways; see Rysiew (2021) for a review. Given that philosophical interest in understanding has only recently achieved critical mass, the more specific research program of a naturalized epistemology of understanding is nascent. We propose to organize much existing work according to the framework in **Figure 1A**. More precisely, philosophical theories of understanding propose how reasoning operates in understanding (see section "Philosophical Theories Propose Reasoning in Understanding (I)"), and these proposals are constrained by explanations and empirical tests found in sciences that study this kind of reasoning (see section "Scientific Studies of Reasoning's Contributions to the Philosophy of Understanding (II)").

Philosophical Theories Propose Reasoning in Understanding (I)

Two kinds of understanding have garnered significant philosophical attention: explanatory understanding (Grimm, 2010, 2014; Khalifa, 2012, 2013a,b, 2017; Greco, 2013; Strevens, 2013; Hills, 2015; Kuorikoski and Ylikoski, 2015; Potochnik, 2017) and objectual understanding (Kvanvig, 2003; Elgin, 2004, 2017; Carter and Gordon, 2014; Kelp, 2015; Baumberger and Brun, 2017; Baumberger, 2019; Dellsén, 2020; Wilkenfeld, 2021). Explanatory understanding involves understanding why or how something is the case. (For terminological convenience, subsequent references to "understanding-why" are elliptical for "understanding-why or –how.") Examples include understanding why Caesar crossed the Rubicon and understanding how babies are made. Objectual understanding is most easily recognized by its grammar: it is the word "understanding" followed immediately by a noun phrase, e.g., understanding Roman history or understanding human reproduction. Depending on the author, the objects of objectual understanding are taken to be subject matters, phenomena, and for some authors (e.g., Wilkenfeld, 2013), physical objects and human behaviors. For instance, it is natural to think of Roman history as a subject matter but somewhat counterintuitive to think of it as a phenomenon. It is more natural to think of, e.g., the unemployment rate in February 2021 as a phenomenon than as a subject matter. Human reproduction, by contrast, can be comfortably glossed as either a subject matter or a phenomenon.

To clarify what they mean by explanatory and objectual understanding, philosophers have disambiguated many other senses of the English word "understanding." Frequently, these senses are briefly mentioned to avoid confusion but are not discussed at length. They are listed in **Table 1**. Scientists may find these distinctions useful when characterizing the kind of understanding they are studying. That said, we will focus on explanatory understanding hereafter. Thus, unless otherwise noted, all subsequent uses of "understanding" refer exclusively to explanatory understanding.

Virtually all philosophers agree that one can possess an accurate explanation without understanding it, e.g., through rote memorization. In cases such as this, philosophers widely agree that the lack of understanding is due to the absence of significant *inferential* or *reasoning* abilities. However, philosophers disagree about *which* inferences characterize understanding. Three broad kinds of reasoning have emerged. First, some focus on the reasoning required to *construct* or *consider* explanatory models (Newman, 2012, 2013, 2015; De Regt, 2017). Second, others focus on the reasoning required to *evaluate* those explanatory models (Khalifa, 2017). On both these views, explanatory models serve as the *conclusions* of the relevant inferences. However, the third and most prominent kind of reasoning discussed takes explanatory information as *premises* of the relevant reasoning—paradigmatically the inferences about how counterfactual changes in the explanatory variable or *explanans* would result in changes to the dependent variable or *explanandum* (Hitchcock and Woodward, 2003; Woodward, 2003; Grimm, 2010, 2014; Bokulich, 2011; Wilkenfeld, 2013; Hills, 2015; Kuorikoski and Ylikoski, 2015; Rice, 2015; Le Bihan, 2016; Potochnik, 2017; Verreault-Julien, 2017). This is frequently referred to as the ability to answer "what-if-things-had-been-different questions." Many of these authors discuss all three of these kinds of reasoning—which we call *explanatory consideration, explanatory evaluation,* and *counterfactual reasoning*—often without explicitly distinguishing them in the ways we have here.

Scientific Studies of Reasoning's Contributions to the Philosophy of Understanding (II)

A naturalized epistemology of understanding begins with the recognition that philosophers do not have a monopoly on studying these kinds of reasoning. Computer scientists, psychologists, and neuroscientists take explanatory and counterfactual reasoning to be important topics of research. Undoubtedly, each discipline has important insights and

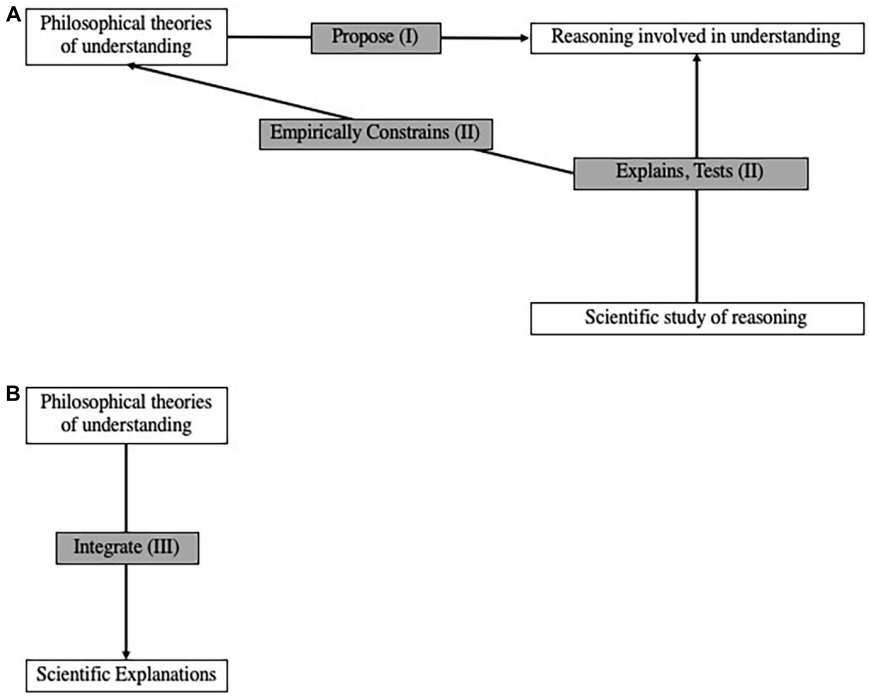


FIGURE 1 | Two ways to integrate philosophical work on understanding with relevant sciences. **(A)** Naturalized epistemology of understanding. **(B)** Understanding-based integration.

contributions. Moreover, these scientific disciplines may raise interesting questions about understanding that are not on the current philosophical agenda.

Cognitive psychological investigations into the nature of explanation and understanding frequently focus on the role of those states in our cognitive lives. To the extent that one can derive a general lesson from this literature, it is probably that both having and seeking explanations aid other crucial cognitive tasks such as prediction, control, and categorization. Developmental psychologists argue that having proper explanations promotes survival, and that at least the sense of understanding evolved to give us an immediate reward for gaining such abilities (Gopnik, 1998). In cognitive psychology, Koslowski et al. (2008) have argued that having an explanation better enables thinkers to incorporate evidence into a causal framework. Lombrozo and collaborators have done extensive empirical work investigating the epistemic advantages and occasional disadvantages of simply being prompted to explain new data. They find that under most normal circumstances trying to seek explanations enables finding richer and more useful patterns (Williams and Lombrozo, 2010). This work also has the interesting implication that the value of explanation and understanding depends on the extent to which there are genuine patterns in the world, with fully patterned worlds granting the most advantages from prompts to explain (ibid.), and more exception-laden worlds providing differential benefits (Kon and Lombrozo, 2019). It has also been demonstrated that attempts to explain can (perhaps counterintuitively) systematically mislead. For example, attempts to explain can lead to miscategorization and inaccurate predictions when there are no real patterns in the data (Williams et al., 2013). Similarly, laypeople can be misguided by the appearance of irrelevant neuroscientific or otherwise reductive explanations (Weisberg et al., 2008; Hopkins et al., 2016). In more theoretical work, Lombrozo (2006) and Lombrozo and Wilkenfeld (2019) consider how different kinds of explanation can lead to understanding that is either more or less tied to specific causal pathways connecting explananda and explanantia vs. understanding focused on how different pathways can lead to the same end result. Thagard (2012) has argued that explanatory

TABLE 1 | Kinds of understanding that philosophers infrequently discuss (Khalifa, 2017, p. 2).

Kind of understanding	Typical complement	Examples
Propositional	That + declarative sentence	I understand that you might not enjoy reading this book.
Broad linguistic	Name of a language	Schatzi understands German.
Narrow linguistic	What + a linguistic expression + means	Schatzi understands what "Ich bin ein Berliner" means.
Procedural	How + infinitive	Miles understands how to play trumpet.
Non-explanatory interrogative	Embedded question that does not seek an explanation as its answer (most who, where, what, and when questions)	I understand who my friends are. I understand where my friends will be going. I understand what my friends are doing. I understand when my friends need a good laugh.

reasoning is key to science's goals both intrinsically and as they contribute to truth and education.

One recent thread in the cognitive science and philosophy of understanding combines insights from information theory and computer science to characterize understanding in terms of data compression. Data compression (Grünwald, 2004) involves the ability to produce large amounts of information from relatively shorter hypotheses and explicitly encoded data sets—in computer science and model-centric physics, there is a burgeoning sense that understanding is tied to pattern recognition and data compression. Petersen (2022)[1] helpfully documents an array of such instances. Li and Vitányi (2008) use compression and explanation almost interchangeably, and at some points even suggest a possible equivalence between compression and the scientific endeavor generally, as in Davies (1990). Tegmark (2014) likewise connects the notion of compression with the explanatory goals of science. Wilkenfeld (2019) translates the importance of compression to good scientific (and non-scientific) understanding into the idiom of contemporary philosophy of science. While part of the inspiration characterizing understanding in terms of compression comes from the traditional "unificationist" philosophical position that understanding involves having to know fewer brute facts (Friedman, 1974) or argument patterns (Kitcher, 1989), the introduction of compression helps evade some objections to unificationist views, such as the fact that such views require explanations to be arguments (Woodward, 2003) and the fact that they allow for understanding *via* unification that no actual human agent can readily use (Humphreys, 1993). [Compression as a marker for intelligence has come under recent criticism (e.g., Chollet, 2019) as only accounting for past data and not future uncertainties; we believe Wilkenfeld's (2019) account evades this criticism by defining the relevant compression partially in terms of usefulness, but defending that claim is beyond the scope of this paper.]

There has also been more direct work on leveraging insights from computer science in order to try to build explanatory schemas and even to utilize those tools to reach conclusions about true explanations. Schank (1986) built a model of computerized explanations in terms of scripts and designed programs to look for the best explanations. Similarly, Thagard (1989, 1992, 2012)—who had previously (Thagard, 1978) done seminal philosophical work on good-making features of explanation and how they should guide theory choice—attempted to automate how computers could use considerations of explanatory coherence to make inferences about what actually occurred.

One underexplored area in the philosophy of understanding and computer science is the extent to which neural nets and deep learning machines can be taken to understand anything. While Turing (1950) famously argued that a machine that could behave sufficiently close to a person could thereby think (and thus, perhaps, understand), many argue that learning algorithms are concerned with prediction *as opposed to* understanding. The most extreme version of this position is Searle's (1980) claim that computers by their nature cannot achieve understanding because it requires semantic capacities when manipulating symbols (i.e., an ability to interpret symbols and operations, and to make further inferences based on those interpretations). Computers at best have merely syntactic capabilities (they can manipulate symbols using sets of instructions, without understanding the meaning of either symbols or operation upon them). However, at the point where deep learning machines have hidden representations (Korb, 2004), can generate new (seemingly theoretical) variables (ibid.), and can be trained to do virtually any task to which computer scientists have set their collective minds (including what looks from the outside like abstract reasoning in IBM's Watson and their Project Debater), it raises vital philosophical questions regarding on what basis we can continue to deny deep learning machines the appellation of "understander."

Elsewhere in cognitive science, early psychological studies of reasoning throughout the 1960s and 1970s focused on deductive reasoning and hypothesis testing (Osman, 2014). A major influence on this trajectory was Piaget's (1952) theory of development, according to which children develop the capacity for hypothetico-deductive reasoning around age 12. The kinds of reasoning studied by psychologists then expanded beyond their logical roots to include more humanistic categories such as moral reasoning (Kohlberg, 1958). The psychology literature offers a rich body of evidence demonstrating how people reason under various conditions. For example, there is ample evidence that performance on reasoning tasks is sensitive to the semantic content of the problem being solved. One interpretation of this phenomenon is that in some contexts, people do not reason by applying content-free inference rules (Cheng and Holyoak, 1985; Cheng et al., 1986; Holyoak and Cheng, 1995). This empirical possibility is of particular interest for philosophers. In virtue of their (sometimes extensive) training in formal logic, philosophers' reasoning practices may be atypical of the broader population. This in turn may bias their intuitions about how "people" or "we" reason in various situations, including when understanding. Another issue raised by sensitivity to semantic content is how reasoning shifts depending on the object of understanding. Although the distinctions explicated by philosophers (e.g., explanatory vs. objectual understanding) are clear enough, it is an open empirical question whether and how reasoning differs *within* these categories depending on the particular object and other contextual factors. As a final example, a further insight from psychology is that people may have multiple modes of reasoning that can be applied to the very same problem. Since Wason and Evans (1974) suggested the idea, dual-process theories have dominated the psychology of reasoning.[2] Although both terminology and precise hypotheses vary significantly among dual-process theories (Evans, 2011, 2012), the basic idea is that one system of reasoning is fast and intuitive, relying on prior knowledge, while another is slow and more cognitively demanding. Supposing two or more systems of reasoning can be deployed in the same situation, one important consideration is how they figure in theories about

[1] Petersen, S. (2022). *Explanation as Compression*.

[2] Though see Osman (2004), Keren and Schul (2009), and Stephens et al. (2018) for examples of criticisms.

the reasoning involved in understanding. To the extent that philosophical accounts are not merely normative but also aim at describing how people actually reason when understanding, psychological studies provide valuable empirical constraints and theoretical considerations.

With the aid of techniques for imaging brains while subjects perform cognitive tasks, neuroscientists have also made great progress in recent decades on identifying regions of the brain involved in reasoning. While that is certainly a worthwhile goal, it may seem tangential to determining the kind of reasoning that characterizes understanding. Here, we suggest two ways in which findings from neuroscience may help with this endeavor. First, neuroscientific evidence can help resolve debates where behavioral data underdetermine which psychological theory is most plausible. More precisely, in cases where competing psychological models of reasoning make the same behavioral predictions, they can be further distinguished by the kinds of neural networks that would implement the processes they hypothesize (Operskalski and Barbey, 2017). For example, Goel et al. (2000) designed a functional magnetic resonance imaging (fMRI) experiment to test the predictions of dual mechanism theory vs. mental model theory. According to the former, people have distinct mechanisms for form- and content-based reasoning, and the latter should recruit language processing structures in the left hemisphere. Mental model theory, by contrast, claims that reasoning essentially involves iconic representations, i.e., non-linguistic representations whose structure corresponds to the structure of whatever they represent (Johnson-Laird, 2010). In early formulations of the theory, it was assumed that different kinds of reasoning problems depend on the same visuo-spatial mechanisms in the right hemisphere (Johnson-Laird, 1995). Goel et al. (2000) tested the theories against one another by giving subjects logically equivalent syllogisms with and without semantic content. As expected, behavioral performance was similar in both conditions. Neither theory predicts significant behavioral differences. Consistent with both theories, the content-free syllogisms engaged spatial processing regions in the right hemisphere. However, syllogisms with semantic content activated a left hemisphere ventral network that includes language processing structures like Broca's area. Unsurprisingly, proponents of mental models have disputed the interpretation of the data (Kroger et al., 2008). We do not take a stance on the issue here. We simply raise the case because it illustrates how neuroscience can contribute to debates between theories of reasoning pitched at the psychological level.

Neuroscientific evidence can also guide the revision of psychological models of understanding and reasoning. The broader point is about cognitive ontology. In the sense we mean here, a cognitive ontology is a set of standardized terms which refer to the entities postulated by a cognitive theory (Janssen et al., 2017). The point of developing a cognitive ontology is to represent the structure of psychological processes and facilitate communication through a shared taxonomy. One role for neuroscience is to inform the construction of cognitive ontologies. Price and Friston (2005), for instance, defend a strong bottom-up approach. In their view, components in a cognitive model (e.g., a model of counterfactual reasoning) should be included or eliminated depending on our knowledge of functional neuroanatomy. Others agree that neuroscience has a crucial role to play in theorizing about cognitive architecture but reject that it has any special authority in this undertaking (Poldrack and Yarkoni, 2016; Sullivan, 2017). We take no position here on how exactly neuroscience should influence the construction of cognitive models and ontologies. Instead, we highlight this important interdisciplinary issue to motivate the potential value of neuroscience for models of understanding and the reasoning involved in it, including those developed by philosophers.

PHILOSOPHICAL THEORIES OF UNDERSTANDING INTEGRATE SCIENTIFIC EXPLANATIONS (III)

Thus, there appear to be ample resources for a naturalized epistemology of understanding, in which explanations and empirical tests from the cognitive sciences empirically constrain philosophical proposals about the kinds of reasoning involved in understanding. However, we offer a second and distinct proposal for how the philosophy of understanding can inform scientific practice: as an account of how different explanations can be integrated (**Figure 1B**).

Such integration is needed when different explanations of a single phenomenon use markedly different vocabularies and concepts. This diversity of explanations is prevalent in several sciences—including the cognitive sciences. To that end, we first present different kinds of explanations frequently found in the cognitive sciences. Whether these different explanations are complements or competitors to each other raises several issues that are simultaneously methodological and philosophical. To address these issues, we then present a novel account of explanatory integration predicated on the idea that explanations are integrated to the extent that they collectively promote understanding. To illustrate the uniqueness of this account, we contrast our account of integration with a prominent alternative in the philosophical literature.

Before proceeding, two caveats are in order. First, although we focus on the cognitive sciences, the account of explanatory integration proposed here is perfectly general. In principle, the same account could be used in domains ranging from particle physics to cultural anthropology. Second, our aim is simply to show that our account of integration enjoys some initial plausibility; a more thoroughgoing defense exceeds the current paper's scope.

A Variety of Scientific Explanations

Puzzles about explanatory integration arise only if there are explanations in need of integration, i.e., explanations whose fit with each other is not immediately obvious. In this section, we provide examples of four kinds of explanations found in the cognitive sciences: mechanistic, computational, topological, and dynamical.

Mechanistic Explanations

Mechanistic explanations are widespread in the cognitive sciences (Bechtel and Richardson, 1993; Machamer et al., 2000; Craver, 2007; Illari and Williamson, 2010; Glennan, 2017; Craver and Tabery, 2019). Despite extensive discussion in the philosophical literature, there is no consensus on the proper characterization of mechanisms or how exactly they figure in mechanistic explanations.[3] For our purposes, we illustrate basic features of mechanistic explanations by focusing on Glennan's (2017, p. 17) minimal conception of mechanisms:

> A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon.

This intentionally broad proposal captures a widely held consensus among philosophers about conditions that are necessary for something to be a mechanism. Where they disagree is about further details, such as the nature and role of causation, regularities, and levels of analysis involved in mechanisms. At a minimum, mechanistic explanations account for the phenomenon to be explained (the *explanandum*) by identifying the organized entities, activities, and interactions responsible for it.

Consider the case of the action potential. A mechanistic explanation of this phenomenon specifies parts such as voltage-gated sodium and potassium channels. It describes how activities of the parts, like influx and efflux of ions through the channels, underlie the rapid changes in membrane potential. It shows how these activities are organized such that they are responsible for the characteristic phases of action potentials. For example, the fact that depolarization precedes hyperpolarization is explained in part by the fact that sodium channels open faster than potassium channels. In short, mechanistic explanations spell out the relevant physical details.

Importantly, not all theoretical achievements in neuroscience are mechanistic explanations. As a point of contrast, compare Hodgkin and Huxley's (1952) groundbreaking model of the action potential. With their mathematical model worked out, they were able to predict properties of action potentials and neatly summarize empirical data from their voltage clamp experiments. However, as Hodgkin and Huxley (1952) explicitly pointed out, their equations lacked a physical basis. There is some disagreement among philosophers about how we should interpret the explanatory merits of the model (Levy, 2014; Craver and Kaplan, 2020; Favela, 2020a), but what is clear is that the Hodgkin and Huxley model is a major achievement that is *not* a mechanistic explanation of the action potential. We return to issues such as these below.

Computational Explanations

Mechanistic explanations are sometimes contrasted with other kinds of explanation. In the philosophical literature, computational explanations are perhaps the most prominent alternative. Computational explanations are frequently considered a subset of *functional explanations*. The latter explain phenomena by appealing to their function and the functional organization of their parts (Fodor, 1968; Cummins, 1975, 1983, 2000). Insofar as computational explanations are distinct from other kinds of functional explanations, it is because the functions to which they appeal involve information processing. Hereafter, we focus on computational explanations.

In computational explanations, a phenomenon is explained in terms of a system performing a computation. A computation involves the processing of input information according to a series of specified operations that results in output information. While many computational explanations describe the object of computation as having representational content, some challenge this as a universal constraint on computational explanations (Piccinini, 2015; Dewhurst, 2018; Fresco and Miłkowski, 2021). We will use "information" broadly, such that we remain silent on this issue. Here, "operations" refer to logical or mathematical manipulations on information such as addition, subtraction, equation (setting a value equal to something), "AND," etc. For example, calculating *n!* involves taking in input *n* and calculating the product of all natural numbers less than or equal to *n* and then outputting said product. Thus, we can explain why pressing "5," "!," "=", in sequence on a calculator results in the display reading "120"; the calculator *computes* the factorial.

More detailed computational explanations of this procedure are possible. For example, the calculator performs this computation by storing *n* and iteratively multiplying the stored variable by one less than the previous iteration from *n* to 1. In this case, the operations being used are equation, multiplication, and subtraction. The information upon which those operations are being performed are the inputted value for *n* and the stored variable for the value of the factorial at that iteration.

Topological Explanations

In topological or "network" explanations, a phenomenon is explained by appeal to graph-theoretic properties. Scientists infer a network's structure from data, and then apply various graph-theoretic algorithms to measure its topological properties. For instance, clustering coefficients measure degrees of interconnectedness among nodes in the same neighborhood. Here, a node's *neighborhood* is defined as the set of nodes to which it is directly connected. An individual node's *local* clustering coefficient is the proportion of edges within its neighborhood divided by the number of edges that could possibly exist between the members of its neighborhood. By contrast, a network's *global* clustering coefficient is the ratio of closed triplets to the total number of triplets in a graph. A triplet of nodes is any three nodes that are connected by at least two edges. An *open* triplet is connected by exactly two edges; a *closed* triplet, by three. Another topological property, average (or "characteristic") path length, measures the mean number of edges needed to connect any two nodes in the network.

In their seminal paper, Watts and Strogatz (1998) applied these concepts to a family of graphs and showed how a network's topological structure determines its dynamics. First, *regular graphs* have both high global clustering coefficients and high average path length. By contrast, *random graphs* have low global

[3] See Craver (2014) for an overview of the latter issue.

clustering coefficients and low average path length. Finally, they introduced a third type of *small-world graph* with high clustering coefficient but low average path length.

Highlighting differences between these three types of graphs yields a powerful explanatory strategy. For example, because regular networks have larger average path lengths than small-world networks, things will "diffuse" throughout the former more slowly than the latter, largely due to the greater number of edges to be traversed. Similarly, because random networks have smaller clustering coefficients than small-world networks, things will also spread throughout the former more slowly than the latter, largely due to sparse interconnections within neighborhoods of nodes. Hence, *ceteris paribus,* propagation/diffusion is faster in small-world networks. This is because the fewer long-range connections between highly interconnected neighborhoods of nodes shorten the distance between neighborhoods of nodes that are otherwise very distant and enables them to behave as if they were first neighbors. For example, Watts and Strogatz showed that the nervous system of *Caenorhabditi elegans* is a small-world network, and subsequent researchers argued that this system's small-world topology explains its relatively efficient information propagation (Latora and Marchiori, 2001; Bullmore and Sporns, 2012).

Dynamical Explanations

In dynamical explanations, phenomena are accounted for using the resources of dynamic systems theory. At root, a system is dynamical if its state space can be described using differential equations, paradigmatically of the following form:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}\left(\boldsymbol{x}(t); \boldsymbol{p}, t\right)$$

Here, x is a vector (often describing the position of the system of interest), $\boldsymbol{f}$ is a function, t is time, and $\boldsymbol{p}$ is a fixed parameter. Thus, the equation describes the evolution of a system over time. In dynamical explanations, these equations are used to show how values of a quantity at a given time and place would uniquely determine the phenomenon of interest, which is typically treated as values of the same quantity at a subsequent time.

For example, consider dynamical explanations of why bimanual coordination—defined roughly as wagging the index fingers of both hands at the same time—is done either in- or anti-phase. Haken et al. (1985) use the following differential equation to model this phenomenon:

$$\frac{d\phi}{dt} = -asin\phi - 2bsin2\phi$$

Here ϕ is relative phase, having a value of either 0° or 180° (representing in- and anti-phase conditions, respectively) and b/a is the coupling ratio inversely related to the oscillations' frequency. The explanation rests on the fact that only the in- and anti-phase oscillations of the index fingers are basins of attraction.

Understanding-Based Integration

Thus far, we have surveyed four different kinds of explanation—mechanistic, computational, topological, and dynamical. Moreover, each seems to have some explanatory power for some phenomena. This raises the question as to how these seemingly disparate kinds of explanation can be integrated. We propose a new account of "understanding-based integration" (UBI) to answer this question. A clear account of understanding is needed if it is to integrate explanations. To that end, we first present Khalifa's (2017) model of understanding. We then extend this account of understanding to provide a framework for explanatory integration.

An Account of Understanding

We highlight two reasons to think that Khalifa's account of understanding is especially promising as a basis for explanatory integration. First, as Khalifa (2019) argues, his is among the most demanding philosophical accounts of understanding. Consequently, it serves as a useful ideal to which scientists should aspire. Second, this ideal is not utopian. This is especially clear with Khalifa's requirement that scientists evaluate their explanations relative to the best available methods and evidence. Indeed, among philosophical accounts of understanding, Khalifa's account is uniquely sensitive to the centrality of hypothesis testing and experimental design in advancing scientific understanding (Khalifa, 2017; Khalifa, in press), and thus makes contact with workaday scientific practices. In this section, we present its three core principles.

Khalifa's first central principle is the *Explanatory Floor*:

> Understanding why *Y* requires possession of a correct explanation of why *Y*.

The Explanatory Floor's underlying intuition is simple. It seems odd to understand why *Y* while lacking a correct answer to the question, "Why *Y*?" For instance, the person who lacks a correct answer to the question "Why do apples fall from trees?" does not understand why apples fall from trees. Since explanations are answers to why-questions, the Explanatory Floor appears platitudinous. Below, we provide further details about correct explanation.

The Explanatory Floor is only one of three principles comprising Khalifa's account and imposes only a necessary condition on understanding. By contrast, the second principle, the *Nexus Principle*, describes how understanding can improve:

> Understanding why *Y* improves in proportion to the amount of correct explanatory information about *Y* (= *Y's* explanatory nexus) in one's possession.

To motivate the Nexus Principle, suppose that one person can correctly identify two causes of a fire, and another person can only identify one of those causes. *Ceteris paribus*, the former understands why the fire occurred better than the latter. Crucially in what follows, however, "correct explanatory information" is not limited to correct explanations. The explanatory nexus also includes the *relationships* between correct explanations. We return to these "inter-explanatory relationships" below.

Furthermore, recall our earlier remark that gaps in understanding arise when one simply has an accurate representation of an explanation (or explanatory nexus) without significant cognitive ability. This leads to the last principle, the *Scientific Knowledge Principle*:

> Understanding why *Y* improves as one's possession of explanatory information about *Y* bears greater resemblance to scientific knowledge of *Y*'s explanatory nexus.

Once again, we may motivate this with a simple example. Consider two agents who possess the same explanatory information that nevertheless differ in understanding because of their abilities to relate that information to relevant theories, models, methods, and observations. The Scientific Knowledge Principle is intended to capture this idea. Khalifa provides a detailed account of scientific knowledge of an explanation:

> An agent *S* has scientific knowledge of why *Y* if and only if there is some *X* such that *S*'s belief that *X explains Y* is the safe result of *S*'s scientific explanatory evaluation (SEEing).

The core notions here are safety and SEEing. Safety is an epistemological concept that requires an agent's belief to not easily have been false given the way in which it was formed (Pritchard, 2009). SEEing then describes the way a belief in an explanation should be formed to promote understanding. SEEing consists of three phases:

1. *Considering* plausible potential explanations of how/why *Y*;
2. *Comparing* those explanations using the best available methods and evidence; and
3. Undertaking *commitments* to these explanations on the basis these comparisons. Paradigmatically, commitment entails that one believes only those plausible potential explanations that are decisive "winners" at the phase of comparison.

Thus, scientific knowledge of an explanation is achieved when one's commitment to an explanation could not easily have been false given the way that one considered and compared that explanation to plausible alternative explanations of the same phenomenon.

Understanding-Based Integration

With our account of understanding in hand, we now argue that it provides a fruitful account of how different explanations, such as the ones discussed above, can be integrated. The Nexus Principle is the key engine of integration. As noted above, this principle states that understanding improves in proportion to the amount of explanatory information possessed. In the cognitive sciences, a multitude of factors explain a single phenomenon. According to the Nexus Principle, understanding improves not only when more of these factors are identified, but when the "inter-explanatory relationships" between these factors are also identified.

One "inter-explanatory relationship" is that of *relative goodness*. Some explanations are *better* than others, even if both are correct. For instance, the presence of oxygen is explanatorily relevant to any fire's occurrence. However, oxygen is rarely judged as the *best* explanation of a fire. Per the Nexus Principle, grasping facts such as these enhances one's understanding. Parallel points apply in the cognitive sciences. For example, it has been observed that mental simulations that involve episodic memory engage the default network significantly more than mental simulations that involve semantic memory (Parikh et al., 2018). Hence, episodic memory better explains cases in which the default network was more active during a mental simulation than does semantic memory.

However, correct explanations can stand in other relations than superiority and inferiority. For instance, the aforementioned explanation involving the default network contributes to a more encompassing computational explanation of counterfactual reasoning involving three core stages of counterfactual thought (Van Hoeck et al., 2015). First, alternative possibilities to the actual course of events are mentally simulated. Second, consequences are inferred from these simulations. Third, adaptive behavior and learning geared toward future planning and problem-solving occurs. The default network figures prominently in the explanation of (at least) the first of these processes (**Figure 2**).

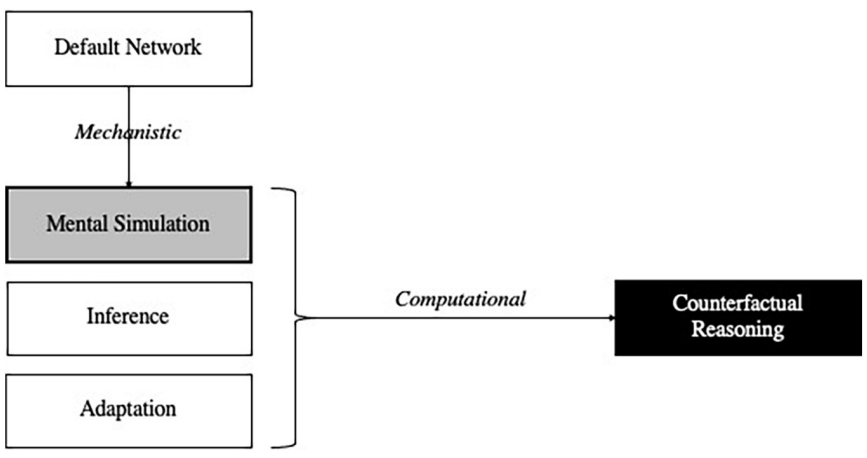


FIGURE 2 | Computational and mechanistic explanations involved in counterfactual reasoning. Mental simulation (gray box) both contributes to the computational explanation of counterfactual reasoning (black box) and is mechanistically explained by the activation of the default network.

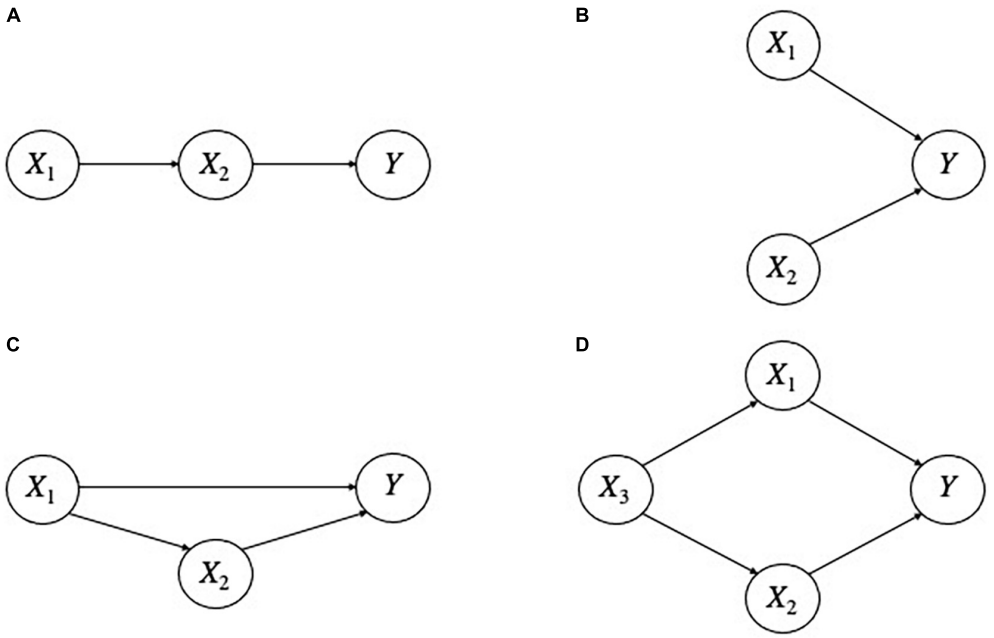


FIGURE 3 | Different inter-explanatory relationships. Letters at the head of an arrow denote phenomena to be explained; those at the tail, factors that do the explaining. Thus, *X1* explains *X2* and *X2* explains *Y* in **(A)**; *X1* and *X2* independently explain *Y* in **(B)**. *X1* explains both *X2* and *Y*, and *X2* also explains *Y* in **(C)**; *X3* explains both *X1* and *X2*, which in turn each explain *Y* in **(D)**.

As this example illustrates, grasping the relationships between different kinds of explanations can advance scientists' understanding. In **Figure 2**, a computational account of mental simulation explains certain aspects of counterfactual reasoning, but mental simulation is then explained mechanistically: the default network consists of parts (e.g., ventral medial prefrontal cortex and posterior cingulate cortex) whose activities and interactions (anatomical connections) are organized so as to be responsible for various phenomena related to mental simulations. Quite plausibly, scientific understanding increases when the relationship between these two explanations is discovered.

Importantly, this is but an instance of an indefinite number of other structures consisting of inter-explanatory relationships (see **Figure 3** for examples). In all of these structures, we assume that for all *i*, X_i is a correct explanation of its respective explanandum. Intuitively, a person who could not distinguish these different explanatory structures would not understand Y as well as someone who did. For instance, a person who knew that X_1 only explains Y through X_2 in **Figure 3A**, or that X_1 and X_2 are independent of each other in **Figure 3B**, or that X_3 is a common explanation or "deep determinant" of both X_1 and X_2 in **Figure 3D**, etc. seems to have a better understanding than a person who did not grasp these relationships. Undoubtedly, explanations can stand in other relationships that figure in the nexus.

Thus, the Nexus Principle provides useful guidelines for how different kinds of explanations should be integrated. Moreover, we have already seen that different kinds of explanations can stand in fruitful inter-explanatory relationships, and that these relationships enhance our understanding. In some cases, we may find that one and the same phenomenon is explained both mechanistically and non-mechanistically, but one of these explanations will be better than another. As noted above, "better than" and "worse than" are also inter-explanatory relationships. So, the Nexus Principle implies that knowing the relative strengths and weaknesses of different explanations enhances understanding.

The Scientific Knowledge Principle also plays a role in UBI. Suppose that X_1 and X_2 are competing explanations of Y. SEEing would largely be achieved when, through empirical testing, X_1 was found to explain significantly more of Y's variance than X_2. This gives scientists grounds for thinking X_1 better explains Y than X_2 and thereby bolsters their understanding of Y. Importantly, SEEing is also how scientists discover other inter-explanatory relationships. An example is the aforementioned study that identified the inter-explanatory relationships between episodic memory, semantic memory, the default network, and mental simulation (Parikh et al., 2018).

Mechanism-Based Integration

Aside from UBI, several other philosophical accounts of explanatory integration in the cognitive sciences are available (Kaplan, 2017; Miłkowski and Hohol, 2020). We provide some preliminary comparisons with the most prominent of these accounts, which we call *mechanism-based integration* (MBI). According to *strong* MBI, all models in the cognitive sciences are explanatory only insofar as they provide information about mechanistic explanations. In response, several critics of MBI—whom we call *pluralists*—have provided examples of putatively non-mechanistic explanations (see **Table 2**). When presented

with putatively non-mechanistic explanations, e.g., of the computational, topological, and dynamical varieties, mechanists (i.e., MBI's proponents) have two strategies available. First, the negative strategy argues that closer scrutiny of the relevant sciences reveals the putatively non-mechanistic explanation to be no explanation at all (Kaplan, 2011; Kaplan and Craver, 2011). The assimilation strategy argues that closer analysis of the relevant sciences reveals the putatively non-mechanistic explanation to be a mechanistic explanation, often of an elliptical nature (Piccinini, 2006, 2015; Piccinini and Craver, 2011; Zednik, 2011; Miłkowski, 2013; Povich, 2015; Hochstein, 2016). Mechanists inclined toward strong MBI frequently use the negative and assimilation strategies in a divide-and-conquer-like manner: the negative strategy applies to some putatively non-mechanistic explanations and the assimilation strategy applies to the rest. However, more prevalent is a *modest* form of MBI that simply applies these strategies to *some* putatively non-mechanistic explanations.

Modest MBI diverges from pluralism on a case-by-case basis. Such cases consist of an explanation where the negative or assimilation strategy seems apt but stands in tension with other considerations that suggest the model is both explanatory and non-mechanistic. On this latter front, several pluralists argue that computational, topological, and dynamical explanations' formal and mathematical properties are not merely abstract representations of mechanisms (Weiskopf, 2011; Serban, 2015; Rusanen and Lappi, 2016; Egan, 2017; Lange, 2017; Chirimuuta, 2018; Darrason, 2018; Huneman, 2018; van Rooij and Baggio, 2021). Others argue that these explanations cannot (Chemero, 2009; Silberstein and Chemero, 2013; Rathkopf, 2018) or need not (Shapiro, 2019) be decomposed into mechanistic components or that they cannot be intervened upon in the same way that mechanisms are intervened upon (Woodward, 2013; Meyer, 2020; Ross, 2020). Some argue that these putatively non-mechanistic explanations are non-mechanistic because they apply to several different kinds of systems that have markedly different mechanistic structures (Chirimuuta, 2014; Ross, 2015). Pluralist challenges specific to different kinds of explanations can also be found (e.g., Kostić, 2018; Kostić and Khalifa, 2022)[4].

In what follows, we will show how UBI is deserving of further consideration because it suggests several plausible alternatives to the assimilation and negative strategies. As such, it contrasts with both strong and modest MBI. While we are partial to pluralism, our discussion here is only meant to point to different ways in which mechanists and pluralists can explore the issues that divide them. Future research would determine whether UBI outperforms MBI.

Assimilation Strategy

According to mechanists' assimilation strategy, many putatively non-mechanistic explanations are in fact elliptical mechanistic explanations or "mechanism sketches" (Piccinini and Craver, 2011; Zednik, 2011; Miłkowski, 2013; Piccinini, 2015; Povich, 2015, in press). Thus, when deploying the assimilation strategy, mechanists take computational, topological, and dynamical models to fall short of a (complete) mechanistic explanation, but to nevertheless provide important information about such mechanistic explanations. Mechanists have proposed two ways that putatively non-mechanistic explanations can provide mechanistic information, and thereby serve as mechanism sketches. First, putatively non-mechanistic explanations can be *heuristics* for discovering mechanistic explanations. Second, putatively non-mechanistic explanations can *constrain* the space of acceptable mechanistic explanations.

An alternative interpretation is possible. The fact that non-mechanistic models assist in the identification of mechanistic explanations does not entail that the former is a species of the latter. Consequently, putatively non-mechanistic explanations can play these two roles with respect to mechanistic explanations without being mere mechanism sketches. In other words, "genuinely" *non-mechanistic* explanations can guide or constrain the discovery of *mechanistic* explanations. Earlier explanatory pluralists (McCauley, 1986, 1996) already anticipated precursors to this idea, but did not tie it explicitly as a response to mechanists' assimilation strategy.

Moreover, this fits comfortably with our account of SEEing and hence with UBI. Heuristics of discovery are naturally seen as advancing SEEing's first stage of considering plausible potential explanations. Similarly, since the goal of SEEing is to identify correct explanations and their relationships, it is a consequence of UBI that different kinds of explanations of the related phenomena constrain each other. For instance, suppose that we have two computational explanations of the same phenomenon, and that the key difference between them is that only the first of these is probable given the best mechanistic explanations of that phenomenon. Then that counts as a reason to treat the first computational explanation as better than the second. Hence, SEEing entails mechanistic explanations can constrain computational explanations.

More generally, UBI can capture the same key inter-explanatory relationships that mechanists prize without assimilating putatively non-mechanistic explanations to mechanistic explanation. Indeed, like many mechanists, UBI suggests that not only do putatively non-mechanistic explanations guide and constrain the discovery of mechanistic explanations, but that the converse is also true. (The next section provides an example of this.) Parity of reasoning entails that mechanistic explanations should thereby be relegated to mere "computational, topological, and dynamical sketches" in these cases, but mechanists must resist this conclusion on pain of contradiction. Since UBI captures these important inter-explanatory relationships without broaching the more controversial question of assimilation, it need not determine which models are mere sketches of adequate explanations. Future research would evaluate whether this is a virtue or a vice.

Negative Strategy

Mechanists' assimilation strategy becomes more plausible than the UBI-inspired alternative if there are good grounds for thinking that the criteria that pluralists use to establish putatively non-mechanistic explanations as genuine explanations

[4]Kostić, D., and Khalifa, K. (2022). *Decoupling Topological Explanation from Mechanisms*.

TABLE 2 | Putatively non-mechanistic explanations discussed by philosophers.

Explanans	Explanandum	Scientific example	Philosophical work discussing example
Computational explanations			
Difference of Gaussians	Stereoscopic vision	Rodieck, 1965; Marr, 1982	Shagrir, 2010; Kaplan, 2011; Kaplan and Craver, 2011*; Bechtel and Shagrir, 2015; Rusanen and Lappi, 2016; Egan, 2017; Shapiro, 2019
Exhaustive search	Recall (memory)	Sternberg, 1969	Shapiro, 2017, 2019
Gain field encoding	Hand–eye coordination	Zipser and Andersen, 1988; Pouget and Sejnowski, 1997; Pouget et al., 2002; Shadmehr and Wise, 2005	Shagrir, 2006*; Kaplan, 2011*; Serban, 2015; Rusanen and Lappi, 2016; Egan, 2017
Geon composition	Object recognition	Hummel and Biederman, 1992	Weiskopf, 2011; Buckner, 2015*; Povich, 2015*
Hybrid computation	Efficiency of brain	Sarpeshkar, 1998	Chirimuuta, 2018
Inhibitory feedback	Normalization	Carandini and Heeger, 2012	Chirimuuta, 2014; Serban, 2015
Internal integration	Eye movement	Seung et al., 2000	Egan, 2017
Line attractor of choice axis, stimuli's selection vector	Context-dependent decision making	Mante et al., 2013	Chirimuuta, 2018
Mapping non-coplanar points to unique rigid configuration	Three-dimensional visual structure of moving objects	Ullman, 1979	Shagrir and Bechtel, 2014*; Egan, 2017
Optimization of spatial and spectral information recovery (Gabor function)	V1 receptive fields	Daugman, 1985	Chirimuuta, 2014, 2018
Similarity of stimulus to stored exemplars	Categorization	Love et al., 2004; Kruschke, 2008	Weiskopf, 2011; Buckner, 2015*; Povich, 2015*
Topological explanations			
Closeness centrality	Speech and tonal processing	Mišić et al., 2018	Kostić, 2020
Mean connectivity	Ictogenicity	Helling et al., 2019	Kostić and Khalifa, 2021
Motif frequency	Functional connectivity	Adachi et al., 2011	Kostić and Khalifa, 2021, 2022 (see text footnote 4)
Navigation efficiency, diffusion efficiency	Efficiency of neuronal communication	Seguin et al., 2019	Kostić, 2020
Network communicability	Cognitive control	Gu et al., 2015	Kostić, 2020
Small-worldness	Information propagation	Watts and Strogatz, 1998	Kostić and Khalifa, 2022 (see text footnote 4)
Dynamical explanations			
Coupling of eye and bodily movements	Onset of motor control	Kelso et al., 1998; Shenoy et al., 2013	Chemero and Silberstein, 2008; Vernazzani, 2019*; Favela, 2020b
Coupling ratio	Bimanual coordination (relative phase)	Haken et al., 1985	Chemero, 2000, 2001; Kaplan and Craver, 2011*; Stepp et al., 2011; Zednik, 2011*; Lamb and Chemero, 2014; Golonka and Wilson, 2019*; Meyer, 2020
Strength of memory trace, salience of target, waiting time, stance	Infant reaching (A-not-B error)	Thelen et al., 2001	Zednik, 2011*; Gervais, 2015; Verdejo, 2015; Venturelli, 2016; van Eck, 2018*; Meyer, 2020; Povich, in press*
Potassium and sodium ion flows	Neural excitability	Hodgkin and Huxley, 1952; FitzHugh, 1961; Nagumo et al., 1962	Craver and Kaplan, 2011*; Kaplan and Bechtel, 2011*; Kaplan and Craver, 2011*; Ross, 2015; Hochstein, 2017*; Favela, 2020a,b

The explanans (first column) is the factor that explains. The explanandum (second column) is the phenomenon to be explained. An asterisk indicates that the author takes the explanation to be mechanistic.

are insufficient. This is the crux of the mechanists' negative strategy. As with the assimilation strategy, we suggest that UBI provides a suggestive foil to the negative strategy.

The negative strategy's key move is to identify a set of non-explanatory models that pluralists' criteria would wrongly label as explanatory. Two kinds of non-explanatory models—how-possibly and phenomenological models—exemplify this mechanist argument. How-possibly models describe factors that *could* but do not *actually* produce the phenomenon to be explained. For instance, most explanations begin as conjectures or untested hypotheses. Those that turn out to be false will be how-possibly explanations. Phenomenological models, which accurately describe or predict the target phenomenon without explaining it, provide a second basis for the negative strategy. Paradigmatically, phenomenological models correctly represent non-explanatory correlations between two or more variables. Mechanists claim that pluralist criteria of explanation will wrongly classify some how-possibly and some phenomenological models as correct explanations. By contrast, since models that accurately represent mechanisms are "how-actually models,"

i.e., models that cite explanatory factors responsible for the phenomenon of interest, MBI appears well-positioned to distinguish correct explanations from how-possibly and phenomenological models.

However, UBI can distinguish correct explanations from how-possibly and phenomenological models. Moreover, it can do so in two distinct ways that do not appeal to mechanisms. First, it can do so on what we call *structural* grounds, i.e., by identifying non-mechanistic criteria of explanation that are sufficient for funding the distinction. It can also defuse the negative strategy on what we call *procedural* grounds, i.e., by showing that the procedures and methods that promote understanding also distinguish correct explanations from these non-explanatory models.

Structural Defenses

We suggest that the following provides a structural defense against the negative strategy:

If *X* correctly explains *Y,* then the following are true:

(1) *Accuracy Condition: X* is an accurate representation, and
(2) *Counterfactual Condition*: Had the objects, processes, etc. represented by *X* been different, then *Y* would have been different.

These are only necessary conditions for correct explanations. They are also sufficient for distinguishing correct explanations from how-possibly and phenomenological models but are likely insufficient for distinguishing correct explanations from every other kind of non-explanatory model. Identifying these other models is a useful avenue for future iterations of the negative strategy and responses thereto.

Situating this within UBI, these conditions are naturally seen as elaborating the Explanatory Floor, which claims that understanding a phenomenon requires possession of a correct explanation. Crucially, mechanists and pluralists alike widely accept these as requirements on correct explanations, though we discuss some exceptions below. Reasons for their widespread acceptance becomes clear with a simple example. Consider a case in which it is hypothesized that taking a certain medication (*X*) explains recovery from an illness (*Y*). If it were discovered that patients had not taken the medication, then this hypothesis would violate the accuracy condition. Intuitively, it would not be a correct explanation, but it would be a how-possibly model.

More generally, how-possibly models are correct explanations *modulo* satisfaction of the accuracy condition. Consequently, pluralists can easily preserve this distinction without appealing to mechanisms; accuracy is sufficient. Just as mechanisms can be either accurately or inaccurately represented, so too can computations, topological structures, and system dynamics be either accurately or inaccurately represented. Similarly, just as inaccurate mechanistic models can be how-possibly models but cannot be correct explanations, so too can inaccurate computational, topological, and dynamical models be how-possibly models but cannot be how-actually models.

Analogously, the counterfactual condition preserves the distinction between correct explanations and phenomenological models. Suppose that our hypothesis about recovery is confounded by the fact that patients' recovery occurred 2 weeks after the first symptoms, and that this is the typical recovery time for anyone with the illness in question, regardless of whether they take medication. Barring extenuating circumstances, e.g., that the patients are immunocompromised, these facts would seem to cast doubt upon the claim that the medication made a difference to their recovery. In other words, they cast doubt on the following counterfactual: had a patient not taken the medication, then that patient would not have recovered when she did. Consequently, the hypothesis about the medication explaining recovery violates the counterfactual condition. Moreover, the hypothesis does not appear to be correct, but would nevertheless describe the patients' situation, i.e., it would be a phenomenological model.

More generally, phenomenological models are correct explanations *modulo* satisfaction of the counterfactual condition. Just as a mechanistic model may accurately identify interacting parts of a system that correlate with but do not explain its behavior, a non-mechanistic model may accurately identify computational processes, topological structures, and dynamical properties of a system that correlate with but do not explain its behavior. In both cases, the counterfactual condition accounts for the models' explanatory shortcomings; no appeal to mechanisms is needed.

Procedural Defenses

Admittedly, structural defenses against the negative strategy are not unique to UBI; other pluralists who are agnostic about UBI have invoked them in different ways. By contrast, our second *procedural* defense against the negative strategy is part and parcel to UBI. Procedural defenses show that the procedures that promote understanding also distinguish correct explanations from how-possibly and phenomenological models.

The Scientific Knowledge Principle characterizes the key procedures that simultaneously promote understanding and distinguish correct explanations from these non-explanatory models. Recall that SEEing consists of three stages: *considering* plausible potential explanations of a phenomenon, *comparing* them using the best available methods, and forming *commitments* to explanatory models based on these comparisons. This provides a procedural defense against the negative strategy. How-possibly and phenomenological models will only be acceptable in the first stage of SEEing: prior to their deficiencies being discovered, they frequently deserve *consideration* as possible explanations of a phenomenon. By contrast, correct explanations must "survive" the remaining stages of SEEing: they must pass certain empirical tests at the stage of comparison such that they are acceptable at the stage of commitment. Indeed, it is often through SEEing that scientists come to distinguish correct explanations from how-possibly and phenomenological models.

Crucially, consideration is most effective when it does not prejudge what makes something genuinely explanatory. This minimizes the possibility of missing out on a fruitful hypothesis. Consequently, both mechanistic and non-mechanistic explanations should be included at this initial stage of SEEing. However, our procedural defense supports pluralism only if some computational, topological, or dynamical explanations are acceptable in light of rigorous explanatory

comparisons. As we see it, this is a strength of our procedural defense, for it uses the empirical resources of our best science to adjudicate debates between mechanists and pluralists that often appear intractable from the philosophical armchair.

Nevertheless, we can point to an important kind of explanatory comparison—which we call *control-and-contrast*—that deserves greater philosophical and scientific attention when considering explanatory integration in the cognitive sciences. Control-and-contrast proceeds as follows. Let X_1 and X_2 be two potential explanations of Y under consideration. Next, run two controlled experiments: one in which the explanatory factors in X_1 are absent but those in X_2 are present and the second in which the explanatory factors in X_1 are present but those in X_2 are absent. If Y is only present in the first experiment, then the pair of experiments suggests that X_2 is a better explanation of Y than X_1. Conversely, if Y is only present in the second experiment, the pair of experiments suggests that X_1 is a better explanation of Y than X_2. If Y is present in both experiments, the experiments are inconclusive. If Y is absent in both experiments, then the experiments suggest that the combination of X_1 and X_2 better explains Y than either X_1 or X_2 does in isolation. Since we suggest that both mechanistic and non-mechanistic explanations should be considered and thereby play the roles of X_1 and X_2, we also suggest that which of these different kinds of explanations is correct for a given phenomenon Y should frequently be determined by control-and-contrast.

In some cases, scientists are only interested in controlling-and-contrasting explanations of the same kind. However, even in these cases, the controls are often best described in terms of other kinds of explanation. For instance, as discussed above, the default mode network mechanistically explains mental simulations involved in episodic memory. By contrast, when mental simulations involve semantic memory, inferior temporal and lateral occipital regions play a more pronounced role (Parikh et al., 2018). Both episodic and semantic memory are functional or computational concepts that can figure as controls in different experiments designed to discover which of these mechanisms explains a particular kind of mental simulation. Less common is controlling-and-contrasting explanations of different kinds. Perhaps this is a lacuna in current research. Alternatively, it may turn out that different kinds of explanation rarely compete and are more amenable to integration in the ways outlined above.

The procedural defense complements the structural defense in two ways. First, not all pluralists accept the accuracy condition. Their motivations for this are twofold. First, given that science is a fallible enterprise, our best explanations today are likely to be refuted. Second, many explanations invoke idealizations, i.e., known inaccuracies that nevertheless enhance understanding. The procedural defense does not require the accuracy condition but can still preserve the distinction between correct explanations and non-explanatory models. Instead, the procedural defense only requires that correct explanations be acceptable on the basis of the best available scientific methods and evidence.

Second, tests such as control-and-contrast regiment the subjunctive conditionals that characterize the counterfactual condition. In evaluating counterfactuals, it is notoriously difficult to identify what must be held constant, what can freely vary without altering the truth-value of the conditional, and what must vary in order to determine the truth-value of the conditional. Our account of explanatory evaluation points to important constraints on this process. Suppose that we are considering two potential explanations X_i and X_j of some phenomenon Y. To compare these models, we will be especially interested in counterfactuals such as, "Had the value of X_i been different (but the value of X_j had remained the same), then the value of Y would have been different," and also, "Had the value of X_i been different (but the value of X_j had remained the same), then the value of Y would have been the same." These are precisely the kinds of counterfactuals that will be empirically supported or refuted by control-and-contrast.

CONCLUSION

Fruitful connections between the philosophy and science of understanding can be forged. In a naturalized epistemology of understanding, philosophical claims about various forms of explanatory and counterfactual reasoning are empirically constrained by scientific tests and explanations. By contrast, in UBI, the philosophy of understanding contributes to the science of understanding by providing broad methodological prescriptions as to how diverse explanations can be woven together. Specifically, UBI includes identification of inter-explanatory relationships, consideration of different kinds of explanations, and evaluation of these explanations using methods such as control-and-contrast. As our suggestions have been of a preliminary character, we hope that future collaborations between philosophers and scientists will advance our understanding of understanding.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors contributed to conception and design of the study. Each author wrote at least one section of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

KK was funded in part by the American Council of Learned Societies' Burkhardt Fellowship. Project Name: Explanation as Inferential Practice. DK was funded by the Radboud Excellence Initiative.

REFERENCES

Adachi, Y., Osada, T., Sporns, O., Watanabe, T., Matsui, T., Miyamoto, K., et al. (2011). Functional connectivity between anatomically unconnected areas is shaped by collective network-level effects in the macaque cortex. *Cereb. Cortex* 22, 1586–1592. doi: 10.1093/cercor/bhr234

Baumberger, C. (2014). Types of understanding: their nature and their relation to knowledge. *Conceptus* 40, 67–88. doi: 10.1515/cpt-2014-0002

Baumberger, C. (2019). Explicating objectual understanding: taking degrees seriously. *J. Gen. Philos. Sci.* 50, 367–388. doi: 10.1007/s10838-019-09474-6

Baumberger, C., and Brun, G. (2017). "Dimensions of objectual understanding," in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, eds S. G. Christoph Baumberger and S. Ammon (London: Routledge), 165–189.

Baumberger, C., Beisbart, C., and Brun, G. (2016). "What is understanding? An overview of recent debates in epistemology and philosophy of science," in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, eds S. R. Grimm, C. Baumberger, and S. Ammon (New York, NY: Routledge), 1–34. doi: 10.1007/978-3-030-38242-1_1

Bechtel, W., and Richardson, R. C. (1993). *Discovering complexity : Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Bechtel, W., and Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Top. Cogn. Sci.* 7, 312–322. doi: 10.1111/tops.12141

Bokulich, A. (2011). How scientific models can explain. *Synthese* 180, 33–45. doi: 10.1007/s11229-009-9565-1

Buckner, C. (2015). Functional kinds: a skeptical look. *Synthese* 192, 3915–3942. doi: 10.1007/s11229-014-0606-z

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336–349. doi: 10.1038/nrn3214

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Carter, J. A., and Gordon, E. C. (2014). Objectual understanding and the value problem. *Am. Philos. Q.* 51, 1–13.

Chemero, A. (2000). Anti-Representationalism and the dynamical stance. *Philos. Sci.* 67, 625–647. doi: 10.1086/392858

Chemero, A. (2001). Dynamical explanation and mental representations. *Trends Cogn. Sci.* 5, 141–142. doi: 10.1016/s1364-6613(00)01627-2

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8367.001.0001

Chemero, A., and Silberstein, M. (2008). After the philosophy of mind: replacing scholasticism with science. *Philos. Sci.* 75, 1–27. doi: 10.1086/587820

Cheng, P. W., and Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cogn. Psychol.* 17, 391–416. doi: 10.1016/0010-0285(85)90014-3

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., and Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cogn. Psychol.* 18, 293–328. doi: 10.1016/0010-0285(86)90002-2

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese* 191, 127–153. doi: 10.1007/s11229-013-0369-y

Chirimuuta, M. (2018). Explanation in computational neuroscience: causal and non-causal. *Br. J. Philos. Sci.* 69, 849–880. doi: 10.1093/bjps/axw034

Chollet, F. (2019). On the measure of intelligence. *arXiv* [Preprint]. arXiv:1911.01547,

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press. doi: 10.1093/acprof:oso/9780199299317.001.0001

Craver, C. F. (2014). "The ontic account of scientific explanation," in *Explanation in the Special Sciences: The Case of Biology and History*, eds I. M. Kaiser, R. O. Scholz, D. Plenge, and A. Hüttemann (Dordrecht: Springer Netherlands), 27–52. doi: 10.1007/978-94-007-7563-3_2

Craver, C. F., and Kaplan, D. M. (2011). "Towards a mechanistic philosophy of neuroscience," in *Continuum Companion to the Philosophy of Science*, eds S. French and J. Saatsi (London: Continuum), 268.

Craver, C. F., and Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *Br. J. Philos. Sci.* 71, 287–319. doi: 10.1093/bjps/axy015

Craver, C. F., and Tabery, J. (2019). "Mechanisms in Science," in *The Stanford Encyclopedia of Philosophy*, Summer 2019 Edn, ed. E. N. Zalta. Available online at: https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/ (accessed August 10, 2021).

Cummins, R. C. (1975). Functional analysis. *J. Philos.* 72, 741–765. doi: 10.2307/2024640

Cummins, R. C. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Cummins, R. C. (2000). ""How does it work?" versus "what are the laws?": Two conceptions of psychological explanation," in *Explanation and Cognition*, eds F. C. Keil and R. A. Wilson (Cambridge, MA: The MIT Press), 117–144.

Darrason, M. (2018). Mechanistic and topological explanations in medicine: the case of medical genetics and network medicine. *Synthese* 195, 147–173. doi: 10.1007/s11229-015-0983-y

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A* 2, 1160–1169. doi: 10.1364/JOSAA.2.001160

Davies, P. C. (1990). "Why is the physical world so comprehensible?," in *Complexity, Entropy and the Physics of Information*, ed. W. Zurek (Boston, MA: Addison-Wesley Publishing Company), 61–70.

De Regt, H. W. (2017). *Understanding Scientific Understanding*. New York, NY: Oxford University Press. doi: 10.1093/oso/9780190652913.001.0001

Dellsén, F. (2020). Beyond explanation: understanding as dependency modelling. *Br. J. Philos. Sci.* 71, 1261–1286. doi: 10.1093/bjps/axy058

Dewhurst, J. (2018). Individuation without representation. *Br. J. Philos. Sci.* 69, 103–116. doi: 10.1093/bjps/axw018

Egan, F. (2017). "Function-theoretic explanation and the search for neural mechanisms," in *Explanation and Integration in Mind and Brain Science*, ed. D. M. Kaplan (Oxford: Oxford University Press), 145–163. doi: 10.1093/oso/9780199685509.003.0007

Elgin, C. Z. (2004). True enough. *Philos. Issues* 14, 113–131. doi: 10.1111/j.1533-6077.2004.00023.x

Elgin, C. Z. (2017). *True Enough*. Cambridge, MA: MIT Press.

Evans, J. S. B. T. (2011). Dual-process theories of reasoning: contemporary issues and developmental applications. *Dev. Rev.* 31, 86–102. doi: 10.1016/j.dr.2011.07.007

Evans, J. S. B. T. (2012). "Dual-process theories of deductive reasoning: facts and fallacies," in *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (New York, NY: Oxford University Press), 115–133. doi: 10.1093/oxfordhb/9780199734689.013.0008

Favela, L. H. (2020a). The dynamical renaissance in neuroscience. *Synthese* 199, 2103–2127. doi: 10.1007/s11229-020-02874-y

Favela, L. H. (2020b). Dynamical systems theory in cognitive science and neuroscience. *Philos. Compass* 15:e12695. doi: 10.1111/phc3.12695

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* 1, 445–466. doi: 10.1016/S0006-3495(61)86902-6

Fodor, J. A. (1968). *Psychological Explanation: An Introduction to the Philosophy Of Psychology*. New York, NY: Random House.

Fresco, N., and Miłkowski, M. (2021). Mechanistic computational individuation without biting the bullet. *Br. J. Philos. Sci.* 72, 431–438. doi: 10.1093/bjps/axz005

Friedman, M. (1974). Explanation and scientific understanding. *J. Philos.* 71, 5–19. doi: 10.2307/2024924

Gervais, R. (2015). Mechanistic and non-mechanistic varieties of dynamical models in cognitive science: explanatory power, understanding, and the 'mere description' worry. *Synthese* 192, 43–66. doi: 10.1007/s11229-014-0548-5

Glennan, S. (2017). *The New Mechanical Philosophy*, 1 Edn. Oxford: Oxford University Press. doi: 10.1093/oso/9780198779711.001.0001

Goel, V., Buchel, C., Frith, C., and Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage* 12, 504–514. doi: 10.1006/nimg.2000.0636

Golonka, S., and Wilson, A. D. (2019). Ecological mechanisms in cognitive science. *Theory Psychol.* 29, 676–696. doi: 10.1177/0959354319877686

Gopnik, A. (1998). Explanation as orgasm. *Minds Mach.* 8, 101–118. doi: 10.1023/A:1008290415597

Gordon, E. C. (2017). "Understanding in Epistemology," in *Internet Encyclopedia of Philosophy*. Available online at: https://iep.utm.edu/understa/ (accessed August 8, 2021).

Greco, J. (2013). "Episteme: knowledge and understanding," in *Virtues and their Vices*, eds K. Timpe and C. A. Boyd (Oxford: Oxford University Press), 285–301. doi: 10.1093/acprof:oso/9780199645541.003.0014

Grimm, S. R. (2010). The goal of understanding. *Stud. Hist. Philos. Sci.* 41, 337–344. doi: 10.1016/j.shpsa.2010.10.006

Grimm, S. R. (2014). "Understanding as knowledge of causes," in *Virtue Epistemology Naturalized*, Vol. 366, ed. A. Fairweather (Dordecht: Springer International Publishing), 329–345. doi: 10.1007/978-3-319-04672-3_19

Grimm, S. R. (2021). "Understanding," in *The Stanford Encyclopedia of Philosophy*, Summer 2021 Edn, ed. E. N. Zalta. Available online at: https://plato.stanford.edu/entries/understanding/ (accessed August 1, 2021).

Grünwald, P. (2004). A tutorial introduction to the minimum description length principle. *arXiv* [Preprint]. math/0406077,

Gu, S., Pasqualetti, F., Cieslak, M., Telesford, Q. K., Yu, A. B., Kahn, A. E., et al. (2015). Controllability of structural brain networks. *Nat. Commun.* 6:8414. doi: 10.1038/ncomms9414

Haken, H., Kelso, J. A. S., and Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biol. Cybernet.* 51, 347–356. doi: 10.1007/BF00336922

Hannon, M. (2021). Recent work in the epistemology of understanding. *Am. Philos. Q.* 58, 269–290. doi: 10.2307/48616060

Helling, R. M., Petkov, G. H., and Kalitzin, S. N. (2019). "Expert system for pharmacological epilepsy treatment prognosis and optimal medication dose prescription: computational model and clinical application," in *Proceedings of the 2nd International Conference on Applications of Intelligent Systems*, (New York, NY: Association for Computing Machinery), doi: 10.1145/3309772.3309775

Hills, A. (2015). Understanding why. *Noûs* 49, 661–688. doi: 10.1111/nous.12092

Hitchcock, C. R., and Woodward, J. (2003). Explanatory generalizations, part II: plumbing explanatory depth. *Noûs* 37, 181–199. doi: 10.1111/1468-0068.00435

Hochstein, E. (2016). One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese* 193, 1387–1407. doi: 10.1007/s11229-015-0844-8

Hochstein, E. (2017). Why one model is never enough: a defense of explanatory holism. *Biol. Philos.* 32, 1105–1125. doi: 10.1007/s10539-017-9595-x

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764

Holyoak, K. J., and Cheng, P. W. (1995). Pragmatic reasoning with a point of view. *Think. Reason.* 1, 289–313. doi: 10.1080/13546789508251504

Hopkins, E. J., Weisberg, D. S., and Taylor, J. C. V. (2016). The seductive allure is a reductive allure: people prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* 155, 67–76. doi: 10.1016/j.cognition.2016.06.011

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Humphreys, P. (1993). Greater unification equals greater understanding? *Analysis* 53, 183–188. doi: 10.2307/3328470

Huneman, P. (2018). Outlines of a theory of structural explanations. *Philos. Stud.* 175, 665–702. doi: 10.1007/s11098-017-0887-4

Illari, P. M., and Williamson, J. (2010). Function and organization: comparing the mechanisms of protein synthesis and natural selection. *Stud. Hist. Philos. Biol. Biomed. Sci.* 41, 279–291. doi: 10.1016/j.shpsc.2010.07.001

Janssen, A., Klein, C., and Slors, M. (2017). What is a cognitive ontology, anyway? *Philos. Explor.* 20, 123–128. doi: 10.1080/13869795.2017.1312496

Johnson-Laird, P. N. (1995). "Mental models, deductive reasoning, and the brain," in *The Cognitive Neurosciences*, ed. M. S. Gazzaniga (Cambridge, MA: MIT Press), 999–1008.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18243–18250. doi: 10.1073/pnas.1012933107

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese* 183, 339–373. doi: 10.1007/s11229-011-9970-0

Kaplan, D. M. (2017). *Explanation and Integration in Mind and Brain Science*, 1st Edn. Oxford: Oxford University Press.

Kaplan, D. M., and Bechtel, W. (2011). Dynamical models: an alternative or complement to mechanistic explanations? *Top. Cogn. Sci.* 3, 438–444. doi: 10.1111/j.1756-8765.2011.01147.x

Kaplan, D. M., and Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philos. Sci.* 78, 601–627. doi: 10.1086/661755

Kelp, C. (2015). Understanding phenomena. *Synthese* 192, 3799–3816. doi: 10.1007/s11229-014-0616-x

Kelso, J. A. S., Fuchs, A., Lancaster, R., Holroyd, T., Cheyne, D., and Weinberg, H. (1998). Dynamic cortical activity in the human brain reveals motor equivalence. *Nature* 392, 814–818. doi: 10.1038/33922

Keren, G., and Schul, Y. (2009). Two is not always better than one: a critical evaluation of two-system theories. *Perspect. Psychol. Sci.* 4, 533–550. doi: 10.1111/j.1745-6924.2009.01164.x

Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philos. Sci.* 79, 15–37. doi: 10.1086/663235

Khalifa, K. (2013a). Is understanding explanatory or objectual? *Synthese* 190, 1153–1171. doi: 10.1007/s11229-011-9886-8

Khalifa, K. (2013b). The role of explanation in understanding. *Br. J. Philos. Sci.* 64, 161–187. doi: 10.1093/bjps/axr057

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.

Khalifa, K. (2019). Is *Verstehen* scientific understanding? *Philos. Soc. Sci.* 49, 282–306. doi: 10.1177/0048393119847104

Khalifa, K. (in press). "Should friends and frenemies of understanding be friends? discussing de Regt," in *Scientific Understanding and Representation: Modeling in the Physical Sciences*, eds K. Khalifa, I. Lawler, and E. Shech (London: Routledge).

Kitcher, P. (1989). "Explanatory unification and the causal structure of the world," in *Scientific Explanation*, Vol. XIII, eds P. Kitcher and W. C. Salmon (Minneapolis, Min: University of Minnesota Press), 410–506.

Kohlberg, L. (1958). *The Development of Modes of Moral Thinking and Choice in the Years 10 to 16.* Ph.D. thesis. Chicago, IL: University of Chicago.

Kon, E., and Lombrozo, T. (2019). "Scientific discovery and the human drive to explain," in *Advances in Experimental Philosophy of Science*, eds D. A. Wilkenfeld and R. Samuels (London: Routledge), 15.

Korb, K. B. (2004). Introduction: machine learning as philosophy of science. *Minds Mach.* 14, 433–440. doi: 10.1023/B:MIND.0000045986.90956.7f

Koslowski, B., Marasia, J., Chelenza, M., and Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cogn. Dev.* 23, 472–487. doi: 10.1016/j.cogdev.2008.09.007

Kostić, D. (2018). The topological realization. *Synthese* 195, 79–98. doi: 10.1007/s11229-016-1248-0

Kostić, D. (2020). General theory of topological explanations and explanatory asymmetry. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375:20190321. doi: 10.1098/rstb.2019.0321

Kostić, D., and Khalifa, K. (2021). The directionality of topological explanations. *Synthese* 199, 14143–14165. doi: 10.1007/s11229-021-03414-y

Kroger, J. K., Nystrom, L. E., Cohen, J. D., and Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Res.* 1243, 86–103. doi: 10.1016/j.brainres.2008.07.128

Kruschke, J. K. (2008). "Models of categorization," in *The Cambridge Handbook of Computational Psychology*, ed. R. Sun (Cambridge: Cambridge University Press), 267—-301.

Kuorikoski, J., and Ylikoski, P. (2015). External representations and scientific understanding. *Synthese* 192, 3817–3837. doi: 10.1007/s11229-014-0591-2

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511498909

Lamb, M., and Chemero, A. (2014). "Structure and application of dynamical models in cognitive science," *in Paper Presented at the 36th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society).

Lange, M. (2017). *Because Without Cause: Non-Causal Explanation in Science and Mathematics*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780190269487.001.0001

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87:198701. doi: 10.1103/PhysRevLett.87.198701

Le Bihan, S. (2016). "Enlightening falsehoods: a modal view of scientific understanding," in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, eds S. R. Grimm, C. Baumberger, and S. Ammon (London: Routledge), 111–136.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *Br. J. Philos. Sci0* 65, 469–492. doi: 10.1093/bjps/axs043

Li, M., and Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and its Applications*, Vol. 3. Cham: Springer. doi: 10.1007/978-0-387-49820-1

Lombrozo, T. (2006). The structure and function of explanations. *Trends Cogn. Sci.* 10, 464–470. doi: 10.1016/j.tics.2006.08.004

Lombrozo, T., and Wilkenfeld, D. (2019). "Mechanistic versus functional understanding," in *Varieties of Understanding*, ed. S. R. Grimm (Oxford: Oxford University Press), 209–230. doi: 10.1093/oso/9780190860974.003.0011

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol. Rev.* 111, 309–332. doi: 10.1037/0033-295x.111.2.309

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philos. Sci.* 67, 1–25.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78–84. doi: 10.1038/nature12742

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman.

McCauley, R. N. (1986). Intertheoretic relations and the future of psychology. *Philos. Sci.* 53, 179–199. doi: 10.1086/289306

McCauley, R. N. (1996). "Explanatory pluralism and the coevolution of theories in science," in *The Churchlands and their Critics*, ed. R. N. McCauley (Hoboken, NJ: Blackwell Publishers), 17–47.

Meyer, R. (2020). The non-mechanistic option: defending dynamical explanation. *Br. J. Philos. Sci.* 71, 959–985. doi: 10.1093/bjps/axy034

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9339.001.0001

Miłkowski, M., and Hohol, M. (2020). Explanations in cognitive science: unification versus pluralism. *Synthese* 199, 1–17. doi: 10.1007/s11229-020-02777-y

Mišić, B., Betzel, R. F., Griffa, A., de Reus, M. A., He, Y., Zuo, X.-N., et al. (2018). Network-based asymmetry of the human auditory system. *Cereb. Cortex* 28, 2655–2664. doi: 10.1093/cercor/bhy101

Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proc. IRE* 50, 2061–2070. doi: 10.1109/JRPROC.1962.288235

Newman, M. (2012). An inferential model of scientific understanding. *Int. Stud. Philos. Sci.* 26, 1–26. doi: 10.1080/02698595.2012.653118

Newman, M. (2013). Refining the inferential model of scientific understanding. *Int. Stud. Philos. Sci.* 27, 173–197. doi: 10.1080/02698595.2013.813253

Newman, M. (2015). Theoretical understanding in science. *Br. J. Philos. Sci.* 68, 571–595. doi: 10.1093/bjps/axv041

Operskalski, J. T., and Barbey, A. K. (2017). "Cognitive neuroscience of causal reasoning," in *The Oxford Handbook of Causal Reasoning*, ed. M. R. Waldmann (New York, NY: Oxford University Press), 217–242.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychon. Bull. Rev.* 11, 988–1010. doi: 10.3758/BF03196730

Osman, M. (2014). "Reasoning research: where was it going? Where is it now? Where will it be going?," in *New Approaches in Reasoning Research*, eds W. De Neys and M. Osman (New York, NY: Psychology Press), 104–123.

Parikh, N., Ruzic, L., Stewart, G. W., Spreng, R. N., and De Brigard, F. (2018). What if? Neural activity underlying semantic and episodic counterfactual thinking. *NeuroImage* 178, 332–345. doi: 10.1016/j.neuroimage.2018.05.053

Piaget, J. (1952). *The Origins of Intelligence in Children* Trans. M. Cook. New York, NY: W. W. Norton & Co. doi: 10.1037/11494-000

Piccinini, G. (2006). Computational explanation in neuroscience. *Synthese* 153, 343–353. doi: 10.1007/s11229-006-9096-y

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.

Piccinini, G., and Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183, 283–311. doi: 10.1007/s11229-011-9898-4

Poldrack, R. A., and Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annu. Rev. Psychol.* 67, 587–612. doi: 10.1146/annurev-psych-122414-033729

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago, IL: The University of Chicago Press.

Pouget, A., and Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* 9, 222–237. doi: 10.1162/jocn.1997.9.2.222

Pouget, A., Deneve, S., and Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nat. Rev. Neurosci.* 3, 741–747. doi: 10.1038/nrn914

Povich, M. (in press). "Mechanistic explanation in psychology," in *The SAGE Handbook of Theoretical Psychology*, eds H. Stam and H. L. De Jong (London: Sage).

Povich, M. (2015). Mechanisms and model-based functional magnetic resonance imaging. *Philos. Sci.* 82, 1035–1046. doi: 10.1086/683438

Price, C. J., and Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275. doi: 10.1080/02643290442000095

Pritchard, D. (2009). Safety-based epistemology: whither now? *J. Philos. Res.* 34, 33–45.

Rathkopf, C. (2018). Network representation and complex systems. *Synthese* 195, 55–78. doi: 10.1007/s11229-015-0726-0

Rice, C. (2015). Moving beyond causes: optimality models and scientific explanation. *Noûs* 49, 589–615. doi: 10.1111/nous.12042

Rodieck, R. W. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vis. Res.* 5, 583–601. doi: 10.1016/0042-6989(65)90033-7

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philos. Sci.* 82, 32–54. doi: 10.1086/679038

Ross, L. N. (2020). Distinguishing topological and causal explanation. *Synthese* 198, 9803–9820. doi: 10.1007/s11229-020-02685-1

Rusanen, A.-M., and Lappi, O. (2016). On computational explanations. *Synthese* 193, 3931–3949. doi: 10.1007/s11229-016-1101-5

Rysiew, P. (2021). "Naturalism in epistemology," in *The Stanford Encyclopedia of Philosophy*, Fall 2021 Edn, ed. E. N. Zalta. Available online at: https://plato.stanford.edu/entries/epistemology-naturalized/ (accessed February 18, 2022).

Sarpeshkar, R. (1998). Analog versus digital: extrapolating from electronics to neurobiology. *Neural Comput.* 10, 1601–1638. doi: 10.1162/089976698300017052

Schank, R. C. (1986). *Explanation Patterns: Understanding Mechanically and Creatively*. Hillsdale, NJ: L. Erlbaum Associates.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Seguin, C., Razi, A., and Zalesky, A. (2019). Inferring neural signalling directionality from undirected structural connectomes. *Nat. Commun.* 10:4289. doi: 10.1038/s41467-019-12201-w

Serban, M. (2015). The scope and limits of a mechanistic view of computational explanation. *Synthese* 192, 3371–3396. doi: 10.1007/s11229-015-0709-1

Seung, H. S., Lee, D. D., Reis, B. Y., and Tank, D. W. (2000). Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron* 26, 259–271. doi: 10.1016/S0896-6273(00)81155-1

Shadmehr, R., and Wise, S. P. (2005). *The Computational Neurobiology of Reaching and Pointing : A Foundation for Motor Learning*. Cambridge: MIT Press.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese* 153, 393–416. doi: 10.1007/s11229-006-9099-8

Shagrir, O. (2010). Marr on computational-level theories. *Philos. Sci.* 77, 477–500. doi: 10.1086/656005

Shagrir, O., and Bechtel, W. (2014). *Marr's Computational Level and Delineating Phenomena*. Oxford: Oxford Unniversity Press.

Shapiro, L. (2017). Mechanism or bust? Explanation in psychology. *Br. J. Philos. Sci.* 68, 1037–1059. doi: 10.1093/bjps/axv062

Shapiro, L. (2019). A tale of two explanatory styles in cognitive psychology. *Theory Psychol.* 29, 719–735. doi: 10.1177/0959354319866921

Shenoy, K. V., Sahani, M., and Churchland, M. M. (2013). Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* 36, 337–359. doi: 10.1146/annurev-neuro-062111-150509

Silberstein, M., and Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philos. Sci.* 80, 958–970. doi: 10.1086/674533

Stephens, R. G., Dunn, J. C., and Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychol. Rev.* 125, 218–244. doi: 10.1037/rev0000088

Stepp, N., Chemero, A., and Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Top. Cogn. Sci.* 3, 425–437. doi: 10.1111/j.1756-8765.2011.01143.x

Sternberg, S. (1969). Memory scanning: mental processes revealed by reaction-time experiments. *Am. Sci.* 57, 421–457.

Strevens, M. (2013). No understanding without explanation. *Stud. Hist. Philos. Sci. A* 44, 510–515.

Sullivan, J. A. (2017). Coordinated pluralism as a means to facilitate integrative taxonomies of cognition. *Philos. Explor.* 20, 129–145. doi: 10.1080/13869795.2017.1312497

Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. New York, NY: Knopf Doubleday Publishing Group.

Thagard, P. (1978). The best explanation: criteria for theory choice. *J. Philos.* 75, 76–92. doi: 10.2307/2025686

Thagard, P. (1989). Explanatory coherence. *Behav. Brain Sci.* 12, 435–502.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton, NJ: Princeton University Press. doi: 10.1515/9780691186672

Thagard, P. (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9218.001.0001

Thelen, E., Schöner, G., Scheier, C., and Smith, L. B. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. *Behav. Brain Sci.* 24, 1–34. doi: 10.1017/s0140525x01003910

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* LIX, 433–460. doi: 10.1093/mind/LIX.236.433

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3877.001.0001

van Eck, D. (2018). Rethinking the explanatory power of dynamical models in cognitive science. *Philos. Psychol.* 31, 1131–1161. doi: 10.1080/09515089.2018.1480755

Van Hoeck, N., Watson, P. D., and Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Front. Hum. Neurosci.* 9:420. doi: 10.3389/fnhum.2015.00420

van Rooij, I., and Baggio, G. (2021). Theory before the test: how to build high-verisimilitude explanatory theories in psychological science. *Perspect. Psychol. Sci.* 16, 682–697. doi: 10.1177/1745691620970604

Venturelli, A. N. (2016). A cautionary contribution to the philosophy of explanation in the cognitive neurosciences. *Minds Mach.* 26, 259–285. doi: 10.1007/s11023-016-9395-0

Verdejo, V. M. (2015). The systematicity challenge to anti-representational dynamicism. *Synthese* 192, 701–722. doi: 10.1007/s11229-014-0597-9

Vernazzani, A. (2019). The structure of sensorimotor explanation. *Synthese* 196, 4527–4553. doi: 10.1007/s11229-017-1664-9

Verreault-Julien, P. (2017). Non-causal understanding with economic models: the case of general equilibrium. *J. Econ. Methodol.* 24, 297–317. doi: 10.1080/1350178X.2017.1335424

Wason, P. C., and Evans, J. S. B. T. (1974). Dual processes in reasoning? *Cognition* 3, 141–154. doi: 10.1016/0010-0277(74)90017-1

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., and Gray, J. R. (2008). The seductive allure of neuroscience explanations. *J. Cogn. Neurosci.* 20, 470–477. doi: 10.1162/jocn.2008.20040

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese* 183, 313–338. doi: 10.1007/s11229-011-9958-9

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese* 190, 997–1016. doi: 10.1007/s11229-011-0055-x

Wilkenfeld, D. A. (2019). Understanding as compression. *Philos. Stud.* 176, 2807–2831. doi: 10.1007/s11098-018-1152-1

Wilkenfeld, D. A. (2021). Objectually understanding informed consent. *Anal. Philos.* 62, 33–56. doi: /10.1111/phib.12173

Williams, J. J., and Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cogn. Sci.* 34, 776–806. doi: 10.1111/j.1551-6709.2010.01113.x

Williams, J. J., Lombrozo, T., and Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. *J. Exp. Psychol. Gen.* 142, 1006–1014. doi: 10.1037/a0030996

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York, NY: Oxford University Press.

Woodward, J. (2013). Mechanistic explanation: its scope and limits. *Proc. Aristotelian Soc. Suppl.* 87, 39–65.

Zednik, C. (2011). The nature of dynamical explanation. *Philos. Sci.* 78, 238–263. doi: 10.1086/659221

Zipser, D., and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331, 679–684. doi: 10.1038/331679a0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 31 March 2022
doi: 10.3389/fnsys.2022.800280

Understanding Is a Process

Leslie M. Blaha[1], Mitchell Abrams[2], Sarah A. Bibyk[1], Claire Bonial[3], Beth M. Hartzler[4], Christopher D. Hsu[3], Sangeet Khemlani[5], Jayde King[1], Robert St. Amant[3], J. Gregory Trafton[5] and Rachel Wong[4]*

[1] 711th Human Performance Wing, U.S. Air Force Research Laboratory, Wright-Patterson Air Force Base, OH, United States, [2] Tufts University, Medford, MA, United States, [3] U.S. Army Combat Capabilities Development Command, Army Research Laboratory, Adelphi, MD, United States, [4] Link Training & Simulation, CAE USA, Arlington, TX, United States, [5] Navy Center for Applied Research in AI, U.S. Naval Research Laboratory, Washington, DC, United States

How do we gauge understanding? Tests of understanding, such as Turing's imitation game, are numerous; yet, attempts to achieve a state of understanding are not satisfactory assessments. Intelligent agents designed to pass one test of understanding often fall short of others. Rather than approaching understanding as a system state, in this paper, we argue that understanding is a process that changes over time and experience. The only window into the process is through the lens of natural language. Usefully, failures of understanding reveal breakdowns in the process. We propose a set of natural language-based probes that can be used to map the degree of understanding a human or intelligent system has achieved through combinations of successes and failures.

Keywords: mutual understanding, common ground, behavioral measurement, human-machine teaming, human-robot interaction, natural language processing, explainable AI, mental models

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Peter Sutor,
University of Maryland, College Park, United States
James Llinas,
University at Buffalo, United States

****Correspondence:***
Leslie M. Blaha
leslie.blaha@us.af.mil

Received: *22 October 2021*
Accepted: *17 January 2022*
Published: *31 March 2022*

Citation:
Blaha LM, Abrams M, Bibyk SA, Bonial C, Hartzler BM, Hsu CD, Khemlani S, King J, St. Amant R, Trafton JG and Wong R (2022) Understanding Is a Process. Front. Syst. Neurosci. 16:800280. doi: 10.3389/fnsys.2022.800280

1. INTRODUCTION

Few would argue with the claim that intelligent behavior in humans and machines depends on *understanding*. Yet, criteria for understanding are elusive. This is because, as this special issue motivates, we know little conclusively about the mechanisms, representations, learning and reasoning that comprise and demonstrate understanding; an ongoing challenge for researchers is to differentiate the unique character of understanding from other cognitive behaviors. A critical step toward establishing a unifying theoretical framework for understanding in both humans and machines is to establish common measures and metrics that elucidate the degree of understanding achieved within candidate frameworks or intelligent systems in a consistent way.

One component of this is clearly articulating what researchers should accept as evidence for understanding, including what constitutes the central tests of a system's ability to understand its input. Hannon (2021) identified a plausible set of criteria for characterizing understanding: understanding is a cognitive achievement, not gained simply by receiving information; understanding comes in degrees; understanding manifests itself through abilities or know-how, especially being able to "grasp" connections. There remains wide disagreement about these basics and even about more fundamental questions, such as whether understanding is a form of knowledge (and thus also subject to questions about the nature of knowledge). But this suggests a single system may exhibit multiple levels of understanding, and these will change over time. Accordingly, the evidence and critical tests should accommodate multiple degrees and adapt over time. Instead of treating understanding as an outcome, it may be more fruitful to consider the question: how does understanding support intelligent behaviors?

In this paper, we argue that understanding is a process, not an outcome. It depends on learning, interpreting, generalizing, and acting upon information. No single test is sufficient for demonstrating that one agent understands another. Indeed, understanding is not a singular type of knowledge (see also, Hannon, 2021). Assessing understanding requires probing the extent of understanding; that is, we need to execute a series of appropriately designed tests that probe the manner and extent to which information has been learned, interpreted, generalized and acted upon. The ability to probe, and therefore demonstrate any degree of, understanding requires natural language.

This paper is organized as follows. Section 1.1 reviews approaches to characterizing understanding from cognitive science and education. Many efforts in these areas attempted to establish comprehensive operational definitions and task-based benchmarks. We identify how agents falling short of desired task performance targets prompts a natural process of probing. Section 2 reviews the closely associated history of major challenge tests for computational intelligence, which place tests of understanding in natural language conversation contexts. Section 3 examines how the challenge of achieving natural language processing in machines has prompted different benchmarks across many levels of meaning representation; both successes and failures at each level illustrate the extent of understanding enabled by each level. Section 4.1 considers the constructive nature of conversation and how humans create mutual understanding through common ground. Despite advances in non-verbal cues for natural interactions (Section 4.2), common ground is a hard challenge for machines, particularly robots. If understanding is a process, then the current inability for machines to understand humans may stem from the inability of machines to engage in the language-dependent process of understanding. Section 5 reviews mental models and theory of mind methods for verbally eliciting knowledge and reasoning from humans. Section 6 reviews recent research on explainable artificial intelligence (XAI), illustrating how machines can make transparent their underlying operations. We synthesize these various approaches from cognitive science, education, natural language understanding, linguistics, verbal protocols, and XAI, to outline a method to craft *probes of understanding* to examine the understanding process. We argue that by establishing such probes in the context of interest, we identify what constitutes evidence for understanding. Thus, we can align the results of probing with the degree to which the desired understanding in humans and machines is achieved and systematically compare hypotheses about the mechanisms underpinning understanding.

1.1. Attempts to Define Understanding

Several broad definitions have been proposed in the cognitive sciences with a goal of establishing a definition that applies to both human and artificial intelligence (AI). For example, Hough and Gluck (2019) recently defined understanding as "The acquisition, organization, and appropriate use of knowledge to produce a response directed toward a goal, when that action is taken with awareness of its perceived purpose" (Hough and Gluck, 2019, p. 23). This is perhaps an updated, more general version of Simon's early definitions developed in his efforts to outline the criteria for software programs capable of understanding. Simon emphasized that understanding is "a relation among a system, one or more bodies of knowledge, and a set of tasks the system is expected to perform" (Simon, 1977, p. 1070). Simon's incorporation of the task or goal for an intelligent system is an extension of Moore and Newell's definition of understanding as a relationship between a system and its appropriate use of knowledge (Moore and Newell, 1974).

Consistently, these definitions emphasize that understanding entails the use of knowledge in pursuit of a task-related goal. Subsequently, the evidence for understanding is then considered to be the ability to successfully perform a target task.

This definition is measurable and achievable within narrowly scoped problems. Narrowly scoped problems include single problem solving tasks (e.g., Towers of Hanoi, demonstrated by the UNDERSTAND program; Simon and Hayes, 1976), or simple information recall in question and answer format (e.g., Siri or similar modern natural-language-based internet search assistants). Throughout the history of AI research, we can find many examples where accomplishing task-related goals has been used to demonstrate success in achieving machine understanding (usually with parallel human demonstrations or baselines).

There is an interesting context in which these early understanding definitions were established. Parallel to the emergence of computing and the computing analogies for cognition in the 1950s and 60s, the first efforts to standardize educational assessment were being published. The first of these, *Taxonomy of Educational Objectives* (Bloom et al., 1956), avoided the use of the term understanding; instead, it emphasized knowledge, comprehension, application, analysis, synthesis, and evaluation as increasingly complex objectives for someone to acquire, interpret, and use information and skills. Revisions and alternatives to this taxonomy replaced use of comprehension with understanding, making it the second level of educational objectives. In the revised *Taxonomy of Educational Objectives*, understanding is currently defined as: "Determining the meaning of instructional messages, including oral, written, and graphic communication" (Krathwohl, 2002, p. 215). This is quite a contrast to the task-oriented definitions in the cognitive sciences. Instead of framing understanding as the successful *use* of knowledge, understanding framed as comprehension emphasizes abilities like interpretation and explanation—abilities that are heavily dependent on natural language communication[1].

However, both the educational taxonomic framing and the task-oriented goal framing of understanding suffer the same pitfall: both frame assessment as pass or fail. An individual is able to pass the test for that level of understanding in the taxonomy or not; an individual can correctly complete the task, or not. Consequently, this pushes the whole construct of understanding to be conceptualized as an intelligent agent's state: it can understand, or it cannot.

A problem with this perspective is that one can pass a test without actually possessing the intended knowledge or

[1]The full list of understanding-related competencies are interpreting, exemplifying, classifying, summarizing, inferring, comparing, explaining (Krathwohl, 2002).

skill, giving an appearance of understanding. When apparent understanding is probed or pushed, perhaps tested in a slightly different context or manner from which the information was learned, the system fails. We see this fragility of performance often for deep neural network classifiers, as evidenced by the discovery of adversarial attacks. In some attacks, very small amounts of noise added to an image can drastically change the confidence of the classifier and switch image class labels (Goodfellow et al., 2018). Very minor changes to the inputs cause sharp increases in classifier errors, indicating that the classifier only had a fragile depth to its representation of the relationship between images and their conceptual-level class assignments. This falls far short of the understanding that developers intended such systems to have.

A danger in chasing the passing of a single test for understanding is that the definition of that test and what it takes to pass become moving targets. Researchers may never agree on a single benchmark against which to measure all claims about mechanisms of understanding. Indeed, Simon (1977) is a microcosm of the dilemma. In a single paper, he lays out at least three full definitions and seven varieties of understanding, because computer programs built to demonstrate sufficient ability for one definition were not sufficient to demonstrate another (see Bobrow and Collins, 1975, for similar examples).

To move our assessments of understanding forward, researchers need to change their perspectives on understanding: namely that understanding is a series of behaviors, not a single outcome.

1.2. The Process of Understanding

We propose that understanding should be conceptualized as a process. Understanding is an ongoing cognitive activity of acquiring, integrating and expressing knowledge according to the task or situation at hand. The process of understanding can amount to an individual's internal reflection on their own knowledge or abilities to accomplish a self-motivated goal; the process by which multiple individuals learn about and communicate with each other while working as a team; and the process of accomplishing team or individual goals. Engaging in the process allows agents to understand themselves, other agents, and external systems or situations. Understanding as a process means that different degrees of understanding may exist in a system, particularly as the tasks or information to be understood are increasingly complex.

Failures of understanding can illustrate breakdowns in the process of understanding. They do this by spotlighting when understanding has not completely enabled success. To determine why an agent failed to understand, failures are usually probed. That is, we find ways to ask why and how thought processes were correct and under what conditions or at what point in reasoning they were not. For example, in educational settings, if a student answers a question incorrectly, they are often asked to explain how they got to the wrong answer (or even to "show their work" to provide teachers with the same information). Cognitive scientists use confusion matrices or patterns of errors to investigate failures of task performance. Both groups try to identify the nature or source of the error, and then try to move toward a state of correcting the error. Hence, probing the failures can result in better understanding. Combined with successes, failures help to map the boundaries or depths of what is and is not understood by an intelligent agent.

2. APPROACH: PROBING FAILURES OF UNDERSTANDING

Assessing understanding as a *process* requires a series of tests that probe a system's successes and failures in different dimensions of understanding. Within AI and Natural Language Processing (NLP), there is a tradition of creating evaluation benchmarks and "challenge" test sets that establish measuring posts of how a system might compare to an ideal, or human-like ability. Perhaps the most well-known of these tests is the "Turing test," proposed by Alan Turing in 1950 to address the question, "Can machines think?" (Turing, 1950). In part due to the difficulties of defining *thinking*, Turing proposed an alternate formulation to probe whether or not machines can exhibit an observable behavior requiring thinking, namely a machine's convincing participation in "the imitation game." In this game, there is a machine, a human participant, and an "interrogator" asking questions of the two parties and viewing written answers to the questions. The interrogator asks questions to ascertain which party is the machine and which is the human. The machine would succeed in this test if it were able to convince the interrogator that it was the human. The Turing test therefore presupposes that the ability to participate in natural conversation evidences intelligent behavior.

Turing hypothesized that a machine would be able to pass his test by the year 2000, and indeed, the Turing test moved from thought experiment to implementation within the Loebner competition starting in 1991—a more limited version of the test in which the interrogator has only 5 minutes to make a determination, and there is a limited set of topics. The first system to pass this limited Turing test selected the topic "whimsical conversation." While fluent, one must question whether such whimsical conversation actually evidences any intelligence (Shieber, 1994). There is enduring fascination with the Turing test that has inspired both a string of philosophical criticisms of it as a litmus test for intelligence as well as alternative tests.

Linguist and philosopher John Searle continued to probe the question "Can computers think?" (Searle, 1984). He concluded that no digital computer can think or "understand" language in particular after posing the "Chinese room experiment." In the Chinese room experiment, he drew a parallel between a person locked in a room manipulating Chinese symbols according to ordering rules (i.e., syntax), but without any knowledge of the actual meaning of these symbols (i.e., semantics), and a computer question-answering system manipulating input symbols designated as questions and returning associated symbols as answers. He concluded that a person in this situation does not "understand" Chinese, and that digital computers are *always* in the Chinese room—while they can manipulate symbols in such a way as to appear to understand language and even

answer questions correctly, they have access only to symbols and syntax, but never the deeper semantics behind those symbols.

Thus, we ask whether or not such evaluations can still have value in their diagnostic ability to pinpoint successes and failures of understanding, where the illusion is broken and we can no longer say that the system functions in practice, regardless of why and how. A system that understands should be able to *articulate* its comprehension and demonstrate its understanding in one or more ways that humans can assess, similar to the ways we have humans demonstrate their comprehension. As a practical matter, this often demands that the system produce responses using natural language. Indeed, we make a strong commitment to the need for processing and responding to natural language: it is only through *natural language probes* that artificial agents can establish their understanding. In the absence of natural language assessments, it may be impossible to establish whether systems are merely symbol-manipulators.

For that reason, we focus on natural language processing as a gateway to understanding in humans and machines. In the following section, we work through attempted assessments of "understanding" in natural language communication, and begin to delineate how we might probe failures in that area to begin to establish benchmarks and metrics for evaluating understanding in a broad variety of systems and tasks.

3. NATURAL LANGUAGE UNDERSTANDING

William James writes, "any number of impressions, from any number of sensory sources, falling simultaneously on a mind WHICH HAS NOT YET EXPERIENCED THEM SEPARATELY, will fuse into a single undivided object for that mind...The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion" (James, 1890, p. 488). Although it has since become debatable how true this is of the human infant brain, this state of blooming buzzing confusion is certainly true for the machine. Similarly, De Saussure writes:

> "Psychologically our thought—apart from its expression in words—is only a shapeless and indistinct mass. Philosophers and linguists have always agreed in recognizing that without the help of signs we would be unable to make a clear-cut, consistent distinction between two ideas. Without language, thought is a vague, uncharted nebula. There are no pre-existing ideas, and nothing is distinct before the appearance of language" (De Saussure, 2011, p. 111).

Again, setting aside debates as to how true this is of human thought, machines must learn how to differentiate sensory input into meaningful bundles—separate categories of the things and events of the world. Furthermore, at least in the domain of the machine's function, they must learn to do so in a way that maps reasonably well to a human's organization of the same sensory input, such that both human and machine can act upon the world in any collaborative task. Because natural language provides a set of labels for many of the discrete categories of the world that humans are familiar with, to come to any kind of understanding between human and machine, the machine must be able to map its own categories and labels to natural language. This amounts to a shared symbolic space between humans and machines, which we propose is critical for establishing understanding and certainly for probing and interrogating a system's level and failures of understanding. It is worth emphasizing that while any shared symbolic space could accomplish this goal, we specifically argue that natural language is the best choice for serving this purpose as the symbolic language most familiar to humans. By "natural language" we are referring to any modality of natural language, in contrast to an artificial, controlled language[2]. Given the fundamental nature of this shared symbolic space to understanding, we discuss in relatively great detail the current landscape of natural language understanding and its evaluation.

3.1. Introduction to Natural Language Understanding

One area of Natural Language Processing (NLP) is referred to as Natural Language Understanding (NLU), a term introduced by Woods (1973), who proposed using English as a query language for a lunar sciences computational system. The motivation for using English as a query language remains relevant today to a variety of applications where NLU components are included. Natural language offers an ease of communication with computational systems, given that people already know, speak, and, as argued by Woods, think, in a natural language. NLU is a higher-order text processing goal, necessarily built upon other NLP components. McCarthy (1990), first published in 1976, proposed what he thought would be the necessary sub-components for achieving NLU:

1. A "parser" that turns English into ANL [Artificial Natural Language].
2. An "understander" that constructs the "facts" from a text in the ANL.
3. Expression of the "general information" about the world that could allow getting the answers to the questions by formal reasoning from the "facts" and the "general information."
    - The "general information" would also contain non-sentence data structures and procedures, but the sentences would tell what goals can be achieved by running the procedures. In this way, we would get the best of the sentential and procedural representations of knowledge.
4. A "problem solver" that could answer the above questions on the basis of the "facts."

Indeed, many NLU approaches introduce a pipeline somewhat like this, including an intermediate, computer-readable semantic representation and knowledge bases that can be used to compare the represented proposition against some real-world knowledge. It is this kind of approach that is also reflected in the discussion

[2]A controlled language could certainly be used to achieve and interrogate understanding in a limited domain, but this places the cognitive burden of communication on the human and precludes efficient generalization to new domains, both of which can be problematic in dynamic and dangerous communication settings.

of semantic processing requirements by Jurafsky and Martin (2009), who indicated that basic requirements include: the truth of the proposition, unambiguous representations drawing upon a specific sense inventory for handling polysemous words and different contexts, as well as the ability to complete disambiguation tasks on the level of both the word and sentence.

3.2. Evaluating Natural Language Understanding

Because the broader goal of NLU is based upon the composition of a variety of lower-level NLP tasks, the question of whether or not a system can successfully "understand" natural language has largely only been addressed first with respect to the particular NLP task at hand (e.g., question-answering), and by evaluating the success of the individual lower-level tasks. Within NLP, these lower-level tasks are most commonly evaluated in the following way:

1. Establish a test set: this is a set of test items, which must be novel items unseen by the system in any training phase. The ground truth result is known, generally by humans establishing this through "annotation" or labeling of text with a set of relevant labels and subsequently comparing annotations for discrepancies to establish an agreed upon "gold standard."
2. Measure the system's ability to reproduce the "gold standard": the most common evaluation metric for this in NLP is an F-score, also referred to as F-measure or F1, which is the harmonic mean of Precision (the number of true positive results divided by the number of all identified positive results) and Recall (the number of true positive results divided by the number of all samples that should have been identified as positive).

For a particular task, accepted baseline and state-of-the-art performance levels are often established through shared tasks, where somewhat different systems with different aims are evaluated on a common test set or suite of test sets. Thus, this is similar to the kind of "challenge" approach, described in Section 2, first established in the Turing test. A good example of a contemporary evaluation suite is The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018), which is a collection of resources for both training and evaluation of various types of NLU tasks. It is intended to be agnostic to the system type. The evaluation suite includes tasks related to sentiment, paraphrase, natural language inference, co-reference, as well as question-answering (many of the challenges present in this evaluation suite parallel the types of probes described in Section 6). Again, system performance on these tasks is often contingent upon the performance of upstream, basic NLP components such as word sense disambiguation and syntactic parsing. In this sense, evaluating and probing the failures of understanding within NLP can be broken down into evaluations of the system's ability to recognize and interpret units of "meaning" at various levels of language, described next.

3.3. Levels of Language Meaning and Understanding

The assumption that a broader NLU task presupposes smaller subtasks reflects assumptions about how and where meaning is encoded in natural language.

3.3.1. Understanding Word Meaning

There is a linguistic tradition that assumes that meaning is compositional—the meaning of a sentence or phrase is made up of the meanings of its individual parts, or word meanings (e.g., Chomsky, 1980). Operating under this assumption, Word Sense Disambiguation (WSD) is a key task for NLU, wherein, given an electronic lexicon or dictionary of word senses, a sense must be assigned to a word in context. For example, the sense of *play* in "She plays the violin" is to perform on an instrument, while "She plays soccer" is to participate in a game. One of the primary challenges of WSD is the selection of an appropriate lexicon, as lexicons can vary greatly in their level of coverage as well as their sense "granularity"—or the number of distinct senses associated with a word. WordNet (Fellbaum, 1998) is probably the most well-known and widely used electronic database of English words with ontological structure. It represents one of the first large-scale efforts to add such structure to a dictionary-like resource. The organization of WordNet was, in part, inspired by work in psycholinguistics investigating how and what type of information is stored in the human mental lexicon (Miller, 1995). WordNet is divided firstly into syntactic categories—nouns, verbs, adjectives and adverbs—and secondly by semantic relations, including synonymy, antonymy, hyponymy (e.g., *tree* is a hypernym of *maple*), and meronymy (part-whole relations). These relations make up a complex network of associations that is both useful for computational linguistics and NLP, and also informative in situating a word's meaning with respect to others.

Although the original English WordNet has been so valuable so as to inspire WordNets in a variety of other languages (e.g., Vossen, 1997), the practical utility of WordNet for WSD tasks has been questioned, as formal evaluations have shown that WordNet's sense inventory is so fine-grained that it is difficult for both humans and systems to tell the difference between senses and apply the appropriate sense label in context. As a response to this, the OntoNotes sense groupings were developed (Pradhan et al., 2007). These can be thought of as a more coarse-grained view of WordNet senses, as these sense groupings were based on WordNet senses that were successively merged into more coarse-grained senses based on the results of measuring inter-annotator agreement (IAA) in tagging of the senses (Duffield et al., 2007). Essentially, where two annotators were consistently able to distinguish between two senses, the distinction was kept. Where annotators were not able to consistently distinguish between two senses, the senses were conflated into one sense. In this way, human IAA establishes the ceiling performance on the task. If humans cannot reliably agree upon the distinctions of an annotation schema, we certainly cannot expect a machine to be able to reproduce those distinctions of manually annotated training and/or test data reliably. Indeed, subsequent systems trained and tested on OntoNotes sense distinctions are able to achieve much better performance on the WSD task, as measured

by F-scores in comparison to a human-annotated gold standard (e.g., Zhong et al., 2008). This has led to OntoNotes becoming a benchmark dataset for training and testing WSD systems.

3.3.2. Understanding Sentence Meaning

Recognizing the meanings of all of the individual words in a sentence, however, does not allow a system to understand the overall meaning of a sentence. We must also enable a system's understanding of *how* meaning is composed, or the semantic relationships between the words. Although there are a variety of established theories as to how to determine and model the semantic relations of a sentence, one dominant assumption widely made in NLP can be summarized Jackendoff's Projection Principle (Jackendoff, 1990), which states that the basic scene denoted by a sentence (i.e., participant roles) derives from the argument structure of the head verb. Verbs structure the relationships between other words of the sentence by designating the "semantic role" that the word plays with respect to the main verb of the sentence. Semantic roles, also called "thematic roles," refer to general classes of participants in a sentence and attempt to define the relation of the participant to the event (which is often expressed by the main verb). For example, in the sentence *Fred gave Maria a book*, *Fred* is the agent of the action, the *book* is the gift, and *Maria* the recipient. The nature of participation in an event for a particular word is often the same, regardless of the syntactic format of the sentence. For example, in *Fred gave a book to Maria*, *Maria* is still the recipient, even though *Maria* is syntactically now an object of a preposition instead of a direct object.

Identifying the semantic roles of the participants is part of the more general task of understanding the semantics of the event, which has certain semantic components regardless of the specific verb used. Whether a speaker talks of *giving, handing,* or *passing,* there is always a transfer of an entity from the giver to a recipient. Grouping verbs with similar semantics allows us to refer to their shared semantic components and participant types. To support a system's ability to recognize and interpret the semantics of a sentence in this way, a variety of resources have been developed wherein human annotators attempt to apply these theories of semantic roles and verb classes to large numbers of English verbs. This annotated data can be used as training and test data for automatic semantic role labeling (SRL), in which a system automatically interprets an the *who, what, where, when, how* of a particular event. SRL resources include the benchmark PropBank (Palmer et al., 2005) and FrameNet (Fillmore et al., 2002) verb lexicons and accompanying annotated corpora, which have been reproduced in a variety of languages.

3.3.3. Understanding Constructional Meaning

NLP has made progress toward recognizing and understanding the meanings of individual words and how those meanings compose to form the meaning of the broader sentence they fall in. Yet, understanding the meaning of a sentence can remain elusive, because there are still other levels of meaning that come into play for a human-like understanding of language. One aspect of this is that, in practice, systems trained on resources that assume the Projection Principle fail to understand sentences where the semantics of participants does not stem from the semantics of the head verb. For example, consider the sentences "She blinked the snow off of her eyelashes," and "We ate our way through New York City." While likely readily understandable to you as the reader, such sentences can be confounding for systems that have been trained to interpret sentence meaning through the lens of the main verb, which is assumed to assign semantic roles to "the snow" and "New York City". This approach leads our systems to expect and likely conclude that snow is something that can be blinked, and a path through New York City is something that can be ingested. Such creative language usages are pushed aside in many linguistic theories as peripheral phenomena of figurative language, unimportant for the broader understanding of language (e.g., Chomsky, 1995). However, the increasing availability of computer-readable corpora has demonstrated the prevalence of these and related phenomena, where the meaning of a sentence is somehow above and beyond the individual word level. In contrast to the Projection Principle, theories of Construction Grammar (e.g., Fillmore, 1988; Goldberg, 1995; Michaelis and Lambrecht, 1996) account for such phenomena. We have begun to see the rise of computational resources (such as the FrameNet "Constructicon"; Fillmore et al., 2012) supporting the recognition and interpretation of "constructions," such as the *caused-motion* and *way-manner* constructions exemplified in the "blink" and "eat" sentences put forth for consideration above.

3.3.4. Understanding Meaning in Conversational Context and Dialogue

Again, even if we add to our system's understanding an interpretation of such constructional meaning beyond the compositional meaning of words, we may be missing implicit information that arises from the broader context of a sentence, from real-world, experiential and cultural knowledge, or from the combination of these factors. This is the broader context involved in dialogue, where language is used in bi-directional communication between speakers or interlocutors. If we would like agents to both understand and potentially communicate about the world around them as another human might, communication *via* natural language dialogue is an appealing candidate. There are significant bodies of research in dialogue systems, which can in turn require computational semantic representations of natural language that attempt to capture all of the levels of meaning described earlier in this section, as well as the recognition of "speech acts," or what someone is attempting to do with a particular utterance beyond its basic content.

Task-oriented spoken dialogue systems, the goal of which is broadly to identify a user's intents and then act upon them to satisfy that intent, have been an active area of research since the early 1990s. Broadly, the architecture of such systems includes (i) automatic speech recognition (ASR) to recognize an utterance in speech and convert this into text, (ii) an NLU component to identify the user's intent, and (iii) a dialogue manager to interact with the user and achieve the intended task (Bangalore et al., 2006). In the earliest of these systems, "understanding" was reduced to the task of detecting a keyword in a user's utterance after the user was prompted with a limited set of permitted options (Wilpon et al., 1990).

Accordingly, the semantic representation within such systems has, in the past, been predefined frames for particular subtasks (e.g., flight inquiry), with slots to be filled (e.g., destination city; Issar and Ward, 1993). In such approaches, the semantic representation was crafted for a specific application, making generalizability to new domains difficult if not impossible. Current approaches still model NLU as a combination of intent and dialogue act classification and slot tagging, but many have begun to incorporate recurrent neural networks (RNNs) and some multi-task learning for both NLU and dialogue state tracking (Chen et al., 2016; Hakkani-Tür et al., 2016), the latter of which allows the system to take advantage of information from the dialogue context to achieve improved NLU. Substantial challenges to these systems include working in domains with intents that have a large number of possible values for each slot and accommodation of out-of-vocabulary slot values (i.e., operating in a domain with a great deal of linguistic variability). Thus, a primary challenge today, as in the past, is representing the meaning of an utterance in a form that can exploit the constraints of a particular domain but also remain portable across domains and robust despite linguistic variability.

There is a long-standing tradition of research in semantic representation within NLP, AI, theoretical linguistics, and philosophy (see Schubert, 2015, for an overview). In this body of research, there are a variety of options that could be used within dialogue systems for NLU. However, for many of these representations, there are no existing automatic "parsers" (which automatically convert language into the representation), limiting their feasibility for larger-scale implementation. Two notable exceptions with a body of research on automatic parsing are combinatory categorial grammar (CCG; Steedman and Baldridge, 2011) and Abstract Meaning Representation (AMR; Banarescu et al., 2013). CCG parsers have already been incorporated in some current dialogue systems (Chai et al., 2014). Although promising, CCG parses closely mirror the input language, so systems making use of CCG parses still face the challenge of a great deal of linguistic variability that can be associated with a single intent. In contrast, AMR abstracts from surface variation; thus, AMR may offer more regular, consistent parses in comparison to CCG. AMR is currently being investigated for use in dialogue systems onboard robots used for search and navigation tasks (Bonial et al., 2019).

To engage in dialogue, an interlocutor must interpret the meaning of a speaker's utterance on at least two levels, as first suggested by Austin (1962): (i) its propositional content, and (ii) its illocutionary force, or the "speech act"—what the speaker is trying to *do* with the utterance in the conversational context. While the aforementioned semantic representations have traditionally sought to represent propositional content, speech act theory has sought to delineate and explicate the relationship between an utterance and its effects on the mental and interactional states of the conversational participants. Speech acts have been used as part of the meaning representation of task-oriented dialogue systems since the 1970s (Bruce, 1975; Cohen and Perrault, 1979; Allen and Perrault, 1980). For a summary of some of the earlier work in this area, see (Traum, 1999). Although the refinement and extension of Austin's (1962) hypothesized speech acts by Searle (1969) remains a canonical work on this topic, there have since been a number of widely used speech act taxonomies that differ from or augment this work, including an ISO standard (Bunt et al., 2012). Nevertheless, these taxonomies often have to be fine-tuned to the domain of interest to be fully useful.

The recognition that meaning representations for dialogue systems need to be expanded to combine different levels of interpretation is growing. For example, Bonial et al. (2020) present Dialogue-AMR, which augments standard AMR, representing the content of an utterance, with speech acts representing illocutionary force. O'Gorman et al. (2018) present a Multi-Sentence AMR corpus (MS-AMR) designed to capture co-reference, implicit roles, and bridging relations. Though not strictly speech acts, the interconnected approach to meaning that this corpus annotates is directly relevant for deducing illocutionary force in a dialogue context.

Although human-robot dialogue systems often leverage a similar architecture to that of the spoken dialogue systems described above, human-robot dialogue introduces the challenge of physically situated dialogue and the necessity for symbol and action grounding, which generally incorporate computer vision. Few systems are tackling all of these challenges at this point (but see Chai et al., 2017). Symbol grounding invokes an additional layer of meaning, as systems must be able to connect a linguistic symbol to a real-world object or event. This requires a challenging combination of both perception of the current environment, as well as real-world knowledge that guide expectations about how to assign sensory input into a category of things grouped under a particular word or label in a given language. In addition to symbol grounding, human-robot dialogue, like human-human dialogue, requires establishing and maintaining "common conversational ground" of the speakers, described further in Section 4.1.

Ontologies have commonly been used for storing, organizing, and deploying the real-world knowledge required for physically situated dialogue systems (as well as other intelligent agents). However, we note that mapping informal concepts into a formal language is a difficult and persistent problem, one in which relatively little progress has been seen. For an example, consider the difficulty of establishing that a machine understands how a box works (Davis, 2011). Even everyday physical concepts that are part of ordinary human conversation, such as near, far, short, friendly, trustworthy, and so forth, are difficult to formalize. A consequence, in part, is that a number of different foundational formalisms (upper ontologies) have been proposed: Basic Formal Ontology (Arp et al., 2015), General Formal Ontology (Herre et al., 2006), Cyc (Matuszek et al., 2006), and others. Despite the challenges, research continues in this area as there are few alternatives that offer any explainability. A research direction that may hold promise is the combination of the value of linguistic and ontological resources with the power of deep learning (e.g., Faruqui et al., 2015).

Overall, the technical landscape of NLU underscores the need for evaluating understanding as a process in which failures can arise at various stages. Probing the success of increasingly complex language understanding tasks allows us to pinpoint and address the limitations of a system's understanding. Although

NLP has established a good model for evaluating systems using suites of benchmark, shared tasks, the evaluations of subtasks within NLP have not been cohesively united to establish clear and measurable evaluations of the most complex tasks that rely on lower levels of understanding. For example, there is little consensus on how to evaluate either "success" or understanding for dialogue systems (see Deriu et al., 2021, for a survey on this topic).

3.4. Generative Language Models

Many of the approaches to different aspects of NLU described thus far have been either semi or fully supervised machine learning, often drawing upon human-annotated training data and possibly some rule-based operations. Recently, NLP has seen the rise of generative language models (GLMs), which constitute a powerful unsupervised approach to various NLP tasks. GLMs produce likely next text based on a context of other text. This process has a surprising number of useful applications, one of which is answering questions about a text passage. This is an application where one may posit that at least certain questions would require understanding of the passage to answer sensibly. One of the most dominant current GLMs is the "Generative Pre-trained Transformer" or GPT. It is a deep neural network with the transformer architecture, trained on a large general text corpus, that generates text as output, given a text prompt.

In contrast with rule-based and/or ontologically-based efforts to provide some knowledge of the meaning behind symbols, recent advances in developing massive pre-trained language models, such as GPT-3 (Brown et al., 2020), have demonstrated successes on a variety of question-answering and inference tasks. GPT-3 illustrates that computers can exploit and deploy knowledge encoded in the text in such a way as to at least broaden and deepen the illusion of understanding language. In part, this success may be attributed to the fact that GPT-3 is trained on huge amounts of text. Thus, whereas the past components that we've looked at are trained on annotated data relating to one or another level of meaning, the broader meaning of entire documents may be implicitly encoded in the GPT-3's training data, giving it a relatively broad "understanding" of meaning in the context of lots and lots of full documents, which can contain a surprising amount of cultural and real-world knowledge.

Nonetheless, GPT-3 has been criticized as "understanding" nothing—criticisms reminiscent of Searle's Chinese Room. Several recent works have set out to pinpoint and classify failures. Drawing inspiration from challenge questions meant to test the strengths and weaknesses of language models like GPT-3 in particular, we suggest the following three dimensions as a starting point for creating probes of a GLM's understanding:

1. **Knowledge Source**: Is the knowledge needed to understand an input contained in information explicitly given to the system, or in the learned world knowledge implicit in the weights acquired during training, or in linguistic knowledge that the system has learned from training?
2. **Knowledge Type**: Is the knowledge needed to understand an input about concrete entities in the world, about events and timelines, or about the contents of the minds of people? Is it about general classes and schemas, or about specific things?
3. **Reasoning Required**: What reasoning abilities are required to understand the input? Can it be answered with analogical, deductive, or inductive logic? Does it require temporal reasoning, reasoning about negation, or meta-reasoning about the motivations of the interlocutor to fully understand?

A recent analysis of the successes and failures of GPT-3 on a question-answering task, involving a carefully curated set of challenge questions, demonstrates that GPT-3 is able to successfully answer questions where the Knowledge Source is explicitly given, and can even answer questions where the knowledge type involves the contents of others' minds and some limited timeline information (Summers-Stay et al., 2021). On the other hand, it is fairly clear that GPT-3 lacks the ability to synthesize and reason about the content it has seen. In particular, GPT-3 has been shown to be unable to perform very simple mathematical operations, even when related to its text prompt, such as using addition or subtraction to determine the age of a person described in a text (Gwern, 2020; Summers-Stay et al., 2021). We suggest that this demonstrates the utility of such challenge sets in probing the failures of understanding and delineating the general areas where a particular system may lack adequate understanding for a particular application or task.

4. DEMONSTRATING AND MAINTAINING SHARED UNDERSTANDING

We now shift from considering natural language understanding, which can be thought of as a largely unidirectional process by which a system interprets and acts upon incoming natural language input, to considerations of how the current level of understanding is *demonstrated* by both humans and machines, and how ongoing shared understanding is maintained. This can be thought of as a bi-directional, dynamic process that may include the initial interpretation of an input, but also the ongoing efforts to subsequently demonstrate that the initial interpretation was or was not successful and then iteratively re-establish that shared understanding is being achieved as communication proceeds.

4.1. Conversation and Common Ground

There is longstanding documentation of the numerous behaviors in which humans engage to cultivate understanding. This includes behaviors designed to establish and maintain what is referred to as the *common ground* (Clark and Wilkes-Gibbs, 1986; Stalnaker, 2002). Common ground is the set of shared beliefs and knowledge that speakers and addressees use to appropriately situate utterances. Information becomes part of the common ground when speakers and addressees demonstrate that they *mutually accept* both the meaning that the speaker intended to convey and that the addressee has understood that meaning. Such information is then said to be *grounded* (Clark and Schaefer, 1987, 1989). The idea that *mutual* acceptance is required for grounding is part of a larger claim that conversation is the *joint* activity of the conversational participants, achieved

through tightly coupled coordination rather than dissociable actions (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark, 1994).

The behaviors that qualify as good "demonstrations" of mutual acceptance are complicated and varied. A behavior that may suffice in one conversational context (e.g., small talk) may be insufficient or inappropriate in another (e.g., defusing a bomb). Grounding behavior also varies as a function of the communication medium (Clark and Brennan, 1991); certain cues for grounding in face-to-face spoken conversation, such as facial expressions or intonation, are unavailable for use in text conversation, though conversational participants can leverage other features of the text medium to ground information (e.g., Potts, 2012; Mills, 2014). In all situations, speakers and addressees must mutually establish an appropriate *grounding criterion* by which to measure whether or not their behaviors demonstrate a reasonable understanding for current purposes (Clark and Wilkes-Gibbs, 1986; Clark, 1994). In some sense, speakers and addressees do not work toward "true" understanding in conversation, but rather toward the belief that there is "sufficient" understanding.

So what are some of the ways in which speakers and addressees contribute to the process of grounding? Speakers often contribute to grounding by working to prevent potential misunderstandings in the first place, such as through "self-repair" of their own utterances; for example, "He called them 'pants' but he meant trousers, like he used the Australian–*the American* word for trousers" where the incorrect "Australian" is immediately corrected to "American" (Schegloff et al., 1977; Clark, 1994). Speakers have been argued to prefer to repair their own utterances, and furthermore initiate those repairs themselves, rather than have their addressee indicate the need for a repair or have the addressee attempt the repair (Schegloff et al., 1977). When prevention of a production error is not possible, speakers may instead warn of possible upcoming understanding difficulties for their addressee through devices such as filled pauses (e.g., "uh" or "um") or other *editing terms* (e.g., the use of "I mean" in an instance such as "We went to the bank–I mean the store"; see Levelt, 1983; Clark, 1994). Speakers cannot always form utterances perfectly, and thus may reformulate their utterances on the fly to improve the likelihood of understanding (Clark and Wilkes-Gibbs, 1986).

Addressees may contribute to grounding through something as simple as continued attention or providing "continuers" (also known as verbal back-channels, such as "mhhm" or "yeah"), or through something as involved as providing an overt indication of understanding through paraphrasing or repeating verbatim what the speaker said (Clark and Schaefer, 1987, 1989). Addressees may also initiate understanding repairs by requesting clarification from the speaker in a form tailored to the nature of their perceived non-understanding (Gonsior et al., 2010). It is through this collaborative effort that conversational participants achieve not only understanding but also the awareness of each other's mutual knowledge required for future conversation.

The legwork that speakers and addressees put into minimizing their *collaborative effort* (even if these contributions sometimes create greater individual effort) not only allow participants to coordinate on their mutual beliefs, but also to develop particular meanings and references as needed in the current task. Such meanings may not extend beyond that task or to new conversational participants (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). These *conceptual pacts* (Brennan and Clark, 1996; Metzing and Brennan, 2003) and language routines (Mills, 2014) present an enormous challenge for human-machine understanding. Creating task-specific meanings (grounded within the task context) is not just served by knowing when and how to deploy collaborative conversational behaviors; arguably such meanings cannot be created without this kind of coordination and negotiation. It is unclear how this form of language innovation and adaptation can be created within human-machine teams until machines possess flexible grounding capabilities, tailored to the medium of communication between the team members.

The fact that human dialogue behaviors are designed to compensate for understanding failures (and such behaviors are arguably like "probing") makes natural language dialogue a fruitful area in which to consider how we might design probes to assess the understanding of artificial systems. However, objectively identifying and quantifying failures of understanding in conversation still presents an enormous challenge. In the absence of overt behavior from the conversational participants themselves, detecting failures requires making assumptions about the mental states of the conversational participants (see Section 5). A distinction is sometimes made between failures of understanding where an addressee is aware of the failure (*non-understanding*) and failures where an addressee is not immediately aware (*misunderstanding*, e.g., Hirst et al., 1994; Weigand, 1999; Gonsior et al., 2010). In the case of non-understanding, addressees take immediate steps to remedy the failure, and therefore there is usually overt evidence in the conversation demonstrating the failure. Clark and Schaefer (1989), for example, identify at least four "states" of understanding in which addressees may believe themselves to be in, and which prompt different kinds of responses to correct the associated failure. The identification and quantification of non-understandings provide a path forward for how we might develop machines that can exhibit similar behaviors (see Gonsior et al., 2010, for one such example). Misunderstandings, on the other hand, must be detected at a later time either by the addressee, the original speaker, or both to be corrected. There may not be overt evidence of a failure at the time the failure occurs. Misunderstandings are ultimately corrected under the assumption that dialogue includes the process of "coming to an understanding," not merely *having* understanding (Weigand, 1999). Further, conversation as a whole is still successful under the assumption that, while at any given moment the conversational participants may be misunderstanding each other, on average understanding is achieved across the entirety of the conversation (Weigand, 1999). The implicit assumption of not only collaboration but *cooperation* within conversation (Grice, 1975) allows humans to progressively and jointly establish understanding. There is much more to be learned about how speakers and addressees balance tolerating some misunderstanding under the assumption that

understanding is being achieved on average, with the need to point out and correct misunderstandings as the conversation progresses. Machines, too, will need to emulate this balance to participate in conversation in a manner that would be perceived as both natural and efficient to a human.

4.2. Perceived Understanding

In some areas of interaction research (e.g., human robot interaction, human-agent interaction), most researchers do not work explicitly on understanding. Most researchers presumably think that understanding *per se* is too difficult a goal to reach during even short-term interaction, so the focus becomes on how to make the robot or agent *appear* as if it were understanding an interaction partner, norms of a situation, or context. We can label these sorts of approaches as *perceived understanding*. Importantly, measures of perceived understanding are usually quite straightforward: preferences and naturalness of the interaction are common metrics.

Most of the work on perceived understanding focuses on cues that the agent or robot can provide that signal that the interaction is progressing. For example, there has been a great deal of work that has shown that appropriate non-verbal communication (eye-gaze, beat gestures, facial expressions) are preferred and considered more natural than either random non-verbal communication or interactions without those cues. Trafton et al. (2008) showed that a robot system that was able to track a conversation non-verbally by looking at the speaker (based on a cognitive model of humans) was perceived as more natural than a system that acted more distracted. Other researchers have also shown that the amount, timing, and location of a robot's gaze can directly impact how much a person wants to interact with the robot (Mutlu et al., 2012; Admoni et al., 2013).

Researchers have also focused on proxemics—the amount of personal space that people maintain around themselves. Takayama and Pantofaru (2009), for example, showed that people became uncomfortable when a robot approached too close to them. Mumm and Mutlu (2011) showed that additional social cues (e.g., head gaze, likability of the robot) interacted with social distance as well. Beat gestures are another form of non-verbal signaling that can be used in interaction. For example, Huang and Mutlu (2013) showed that an agent that provides beat gestures while talking is perceived as more natural. Nods by agents and robots have also been shown to improve interaction and the naturalness of the system (Sidner et al., 2006; Arimoto et al., 2014).

Machines that demonstrate understanding of humans (whether they truly possess such understanding or not) still clearly represent an important benchmark toward creating machines that humans in turn feel they can understand (see Section 6 for further discussion on XAI). For humans to feel that they can probe the understanding of machines in the same manner as human conversational partners, machines must possess the propensity to engage collaboratively and cooperatively with humans in achieving understanding, rather than focusing on the unilateral direction of the machine understanding the human. One possible path toward unqualified human-machine partnership and understanding may require stepping back to better assess the foundations of most human collaborations. Once a common interest or goal has been realized, the next steps are likely to include considering the expectations and thought process of the other, and recognizing how these may differ from your own.

5. APPROACHING UNDERSTANDING FROM MENTAL MODELS AND THEORY OF MIND

A central part of the process of understanding a phenomenon is to build a model of it, i.e., a representation of its salient and functional components. Models may look very different from the phenomenon itself. For example, a watch serves as a model of the rotation of the earth. In cognitive science, human factors, and computer science, researchers agree that humans build models mentally to understand situations or other agents. When a set of individuals build mental models that overlap with one another, they are able to communicate efficiently and, as a consequence, carry out tasks that demonstrate shared understanding.

In this section, we will review the various mental model concepts and measurement methods, as well as theory of mind indicators of inferences about the state of other agents, and examine how each method may help provide insight on understanding in humans and AI systems.

5.1. Mental Model Definitions and Theory

There are multiple perspectives on the definition of mental models. Johnson-Laird (1983) defines mental models as small-scale mental simulations of the world we develop to enable reasoning about the environment around us. Gentner and Stevens (2014) adds that mental models are representations users develop of an environment, situation, or other agent. These models are developed through interaction with a system as well as the user's inferences about the situation or system behaviors. Mental models can be influenced by users' previous experiences such as their exposure to technology and similar systems (Gentner and Stevens, 2014). Most researchers and scientists agree on the ways mental models support intelligent behaviors:

> "Mental models are the mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states" (Rouse and Morris, 1986, p. 3).

Shared mental models are similar; however, shared mental models are the common representations humans have about the functioning, states, and future states of systems. Shared mental models are usually investigated at a team level where the "system" being represented can be a system a team uses together or the "team" itself and its members (Cannon-Bowers et al., 1993; Kennedy et al., 2008; Jonker et al., 2010). Previous research suggests improved mental models and shared mental models are positively correlated with improved individual and team performance. Effective mental models have also been

linked to better situational awareness of a system and improved metacognition (Salas et al., 1994; Scielzo et al., 2004).

5.2. Ways of Measuring Mental Models

There are various methods for mental model elicitation, and each measurement specifically addresses certain aspects within mental model theory. *Think Aloud* methods are one set of mental model elicitation techniques. This method encourages participants to verbally express their thought process about a situation or while completing a task. Participants are guided through the steps of describing their cognitive processes explicitly, often through verbal protocols such as think-out-loud challenges, prospect, and task reflection (Hoffman et al., 2018). One example of this technique is the Think Aloud Problem Solving Task. During this process participants verbally describe their thought process as they complete a task. This method helps to provide insight into how participants frame problems and the steps they take to solve an issue. As participants explain their thoughts, experimenters assess how participants conceptualize a system or issue (Hoffman et al., 2018). Task reflection is a similar technique, where experimenters probe participants post-task about their thought process for completing the task. These methods (e.g., structured interviews, self-explanation task, prediction task) primarily focus on the user's overall representation of the system, approach toward problem-solving, and task reflection/execution (Hoffman et al., 2018).

Another set of elicitation methods draw on how participants understand concepts and their relations to each other, typifying the various components and creating groups of similar factors. Examples of these methods include card sorting, pathfinder, and familiarity ratings. During card sorting and pathfinder methods, participants group similar concepts together and rate how similar each concept is with each other (Hoffman et al., 2018). This measure can help participants schematically represent their conceptualization of a system, its components, and the relationships among items. Diagramming is another mental model elicitation technique, where users can freely draw a pictorial representation of their cognitive process, system, or events (Hoffman et al., 2018). This method can help eliminate the bias of the experimenter on how the user pictorially represents their mental model arrays and may capture new relationships and spatial orientations of concepts.

5.3. Probing Mental Model Failures

Elicitation approaches can easily help researchers identify failures of understanding and gaps in someone's mental model of a system. While conducting these elicitation methods, scientists are able to identify where there is a gap in understanding and the nature of the individual's failed understanding, providing rich information to equip scientists to repair where the misunderstanding occurred. For instance, a novice mechanic could be asked to diagram the layout of an engine and to *Think Out Loud* the process they would take to complete an engine repair. With the assistance of a subject matter expert, scientists can easily determine whether the participant is lacking knowledge of the schematic layout of the engine or if the mechanic is still unfamiliar with the repair process.

While these methods seem to be very insightful for measuring users' representations of systems, these methods of mental model measurement may not have the ability to capture the entirety of understanding, especially when measuring a human's understanding of another human being. Previous research outlines the variability in mental models. Gentner and Stevens (2014) suggest that mental models are unstable mental representations. Additionally, mental models are often incomplete and lack firm boundaries. This is especially true when measuring one's mental model of an unceasingly evolving system. As teammates and humans continue to interact and gain more information about each other, mental models change. One teammate's mental model of their fellow team member may change as they continue to work together; experiences help team members learn more about their teammates' experiences and knowledge. Additionally, as a team faces new challenges together, new knowledge is built and then processed, changing each member's mental model of the world around them, their task, and their teammates. Mental model measurements also fail to capture attitudes and emotional relations between human and human mental models; these aspects are key and crucial to how mental models of teammates are used when completing tasks and relating with one another. We theorize that while mental model measurements may provide effective probing mechanisms for a user's understanding of a system, it may lack the robustness to comprehensively measure and capture a human's "understanding" of another human. Therefore, leading us to believe that understanding may be a bit more intricate and sophisticated than a mental model representation, especially when the subject of the mental model is complex and continually evolving.

5.4. Un-testable Theories in Theory of Mind

The shallowness of these representations is also evident for most measures of theory of mind (ToM), an extension of mental models in that both consider the knowledge or awareness of someone else, yet takes the additional step appreciating how that framework may differ from your own experience. This ability to recognize another's mental state as different from one's own is most commonly operationally measured through counterfactual reasoning or false belief (e.g., Sally-Anne task; Baron-Cohen et al., 1985, though ToM has been demonstrated across a host of situations), such completion of another's failed action and recognizing another's capacity to have concurrent yet conflicting desires (Beaudoin et al., 2020). This ability to hypothesize about the knowledge and intentions of another agent, whether living or synthetic, develops at an early age (Beaudoin et al., 2020) and is a valuable skill for social interactions and effective teaming. In human-human teams, ToM is considered critical to ensuring constructive planning and exchanges toward accomplishing a task, whereas the benefit in human-machine interactions is somewhat more ill-defined yet still seen as important (Benninghoff et al., 2013; Winfield, 2018).

As noted with mental models, numerous measures have been developed to evaluate an individual's capacity for ToM, yet the overwhelming majority of these are only sensitive to developmental stages and clinical populations (Beaudoin

et al., 2020). Such tasks typically ask participants to adopt the perspective of a character in the story who has incomplete knowledge of the situation, then infer how that character is likely to respond. Moreover, most such tasks rely on drawings or situational schematics to describe a third-person account of a fictional scenario, similar to mental model elicitation approaches. However, imaging studies indicate such experiences fail to elicit the same neural response evident for actual social interactions, suggesting participants do not perceive these narratives in a way that accurately replicates personal interactions (Byom and Mutlu, 2013).

More interactive methods have been used, such as Meltzoff's behavioral re-enactment study (Meltzoff, 1995) which demonstrated that 18-month-olds were able to correctly interpret and complete target actions the experimenter initiated but did not finish. Though these results are compelling, paralleling the Chinese room experiment, it is impossible to conclude whether the toddlers had actually inferred the experimenter's intention, or were simply imitating an adult, behavior common for that age group (Jones, 2009). Additionally, studies involving neuro-typical adults have evaluated both observed behaviors in a communication game (Keysar et al., 2003) as well as self-reported experiences during daily activities (Bryant et al., 2013), and concluded that adults, although capable of forming a ToM, actually used the skill very rarely during real-world interactions.

In light of these findings for ToM, as well as those related to mental models outlined previously, the ability to generate any type of insight into the thoughts and perceptions of others is no doubt beneficial, both in casual and teaming environments. Indeed, the capacity to form mental models and ToM is particularly useful across a wide variety of inter-personal situations, such as supporting effective negotiations (de Weerd et al., 2017), and learning or adopting more sophisticated societal norms for ethics and morality (Leslie et al., 2006). It is important to note however that both mental models and ToM are thought to be beneficial precisely because they may help to avoid misunderstandings and failures in collaboration, yet implementation of the metrics discussed above offers little in the way of ensuring two agents have a shared understanding. Thus, members of a team, either human and synthetic, may adequately demonstrate these skills of social cognition, but this should not be viewed as a proxy for ensuring all teammates have a shared understanding.

6. IMPLIED DEFINITIONS OF UNDERSTANDING: EXPLAINABLE AI

One might plausibly think that artificial intelligence is at its core the study of systems that understand. (McDermott, 1976, p. 4) notes a temptation to assume away the challenge, however: "If a researcher tries to write an "understanding" program, it isn't because he has thought of a better way of implementing this well-understood task, but because he thinks he can come closer to writing the *first* implementation." In the intervening half-century we have not yet seen that first implementation.

Relatively little research in AI explicitly addresses understanding in computer systems or its assessment (Thórisson et al., 2016). Simon and Eisenstadt (2000) are an exception. They propose that artificial understanding be treated no differently from human understanding, with conventional psychological tests being applied. They further propose that, in contrast to human testing, we have direct access to an AI system's internal program structures and memory, which may provide evidence for or against understanding: for example, whether a necessary perceptual discrimination is present, or whether a given capability has been learned or was pre-programmed.

Páez (2019), writing about systems that explain their own behavior, is also an exception. Páez holds that explanation should not be the goal for explainable AI (XAI) systems—rather, "a pragmatic and naturalistic account of understanding" should be the focus of the field. Such an account is currently lacking. Research in XAI offers promising hints about understanding, however, which we pursue in the remainder of this section. Our coverage of XAI, to include intelligible systems (Páez, 2019; Weld and Bansal, 2019), transparent systems (Castelvecchi, 2016), and related categories, will be selective. More comprehensive resources are Confalonieri et al. (2021)'s history, Vilone and Longo (2020)'s systematic review, and Mueller et al. (2019)'s meta-review and bibliography.

As a preliminary, note that it is common to probe a person's understanding of some phenomenon by requesting explanations, as in the verbal protocols discussed in Section 5; every schoolchild is familiar with "Explain this…" test questions. This is a form of abduction: we use the requests for explanations as probes, with responses providing evidence for or against specific forms of understanding. Now consider an XAI system, or even all XAI systems. We can translate the implemented explanations and explanatory processes into probes. Because we focus on probing for failure, we do not need to attribute understanding to these systems; rather, each failed probe is interpreted as demonstrating a lack of understanding.

By "translating" an explanation into a probe, we mean that an explanation is typically a carefully structured account that contains different kinds of information. Each is a potential type of probe. We outline major categories below. We label each category, describe representative types of probes found in the literature, and give an example template for a probe. For simplicity, assume that the target phenomenon to be explained (and implicitly, understood) is a behavior y of a given system, and that a probe is of the properties of some set of measurements X of the system or of the environment (which the system may be able to observe or change).

Relevant information. In a symbolic reasoning system, a discrete item of information may be relevant because it is required to make a potential inference (Buchanan and Shortliffe, 1984) or to enable a step in a plan (Fox et al., 2017; Chakraborti et al., 2020). Image classification systems process information in which sets of items may be relevant rather than individual items (e.g., edges or patches rather than pixels). A well-known non-XAI example is Pomerleau (1992)'s discovery that ALVINN, an early autonomous vehicle, had learned to use the amount of grass visible alongside the road as a surrogate for the road's curvature

when it needed to turn, causing unexpected behavior in non-grassy settings. Comparable examples are now commonplace in XAI systems for deep learning (Xu et al., 2019). A simple probe might take the form, "Does *y* vary predictably with different values of *X*?" where *X* may represent different sets of measured variables.

Relevant distinctions. In some cases, in particular for systems that deal with non-discrete data, distinctions are needed even to define relevant features. These include distinguishing features in image classification, which may be highlighted as patches, colored overlays, saliency maps, etc. (Nourani et al., 2019; Xu et al., 2019), as briefly discussed in Section 6.1; differences between term frequencies in text information retrieval (Hearst, 1995); and threshold values or functions on continuous data (Buchanan and Shortliffe, 1984). A probe might take the form, for *X* known to be relevant information, "Does *y* vary predictably with different possible values of *X* only within a specific range of *X*? What is that range?"

Relevant relationships. Treating relationships as a separate category from items of information is largely arbitrary, in that the relationships themselves are information, as are properties of relationships. The distinction can be convenient for discussion, however. Relationships can be relevant in different ways. In explainable AI planning (Fox et al., 2017), different types of relationships between actions may be relevant: temporal ordering; "causal" relationships, i.e., in a causal-link planning sense (Young et al., 1994); the absence of predecessor actions needed for a given action; etc. A naïve Bayes classifier is considered highly explainable in that it explicitly identifies input variables relevant to the output classification variable (Kononenko, 2001). More generally, a Bayes network may be interpreted as a causal model, in which the existence of individual links is relevant: *X* may be the set of causes for *y*, for example. A simple probe might take the form, for different *X*, "If *X* were constant, at different possible values, would *y* vary predictably all of the time? None of the time?"

The probes above address "local" aspects of a phenomenon. Further, there is an emphasis on prediction, though predictive accuracy is not generally considered sufficient for explanation or understanding. We can also consider the more global structure and content of an explanation as evidence for understanding.

Counterfactuals. An account of what would happen under different conditions is important in explanation (e.g., Fox et al., 2017; Korpan and Epstein, 2018) in part because it can be evidence for understanding in terms of causation. Again, these are the central probes and explanations sought for theory of mind assessments. Some of the example probes expressed above have this flavor, e.g., "If the values of *X* were such and such, what would happen to *y*?"

Generalizations and abstractions. If we are interested in *y* under many different values of *X*, we can think of our goal as mapping out the policy that governs the system's behavior. A large number of individual samples may be adequate, but a more concise generalization may be possible, ideally one that applies to values of *X* and *y* not yet observed. This is a goal of ambitious work by Thórisson et al. (2016) in the area of artificial general intelligence. They directly define understanding of a phenomenon Φ as a set of models capable of predicting, explaining, recreating, and achieving goals with respect to Φ.

Analogical cases. Relatedly, if a phenomenon is understood in one domain, it may be possible to transfer that understanding to a new domain. For example, in robot behaviors, navigating to a given location and reaching out to grasp a target object generally depend on different control mechanisms and environment observations. Nevertheless the concept of "blockage of the path" is a generalization for some kinds of failure (St. Amant et al., 2019); each is a plausible analogy for the other.

For all of these types of probes, we require some ground truth against which we can compare a probe's output. Is a system capable of evaluating relevance, making appropriate distinctions, identifying related entities with respect to some phenomenon, in particular its own behavior? Can it extrapolate, answer "What if?" questions, explain how unlike situations actually share some underlying similarities? As we walk through a set of probes, we accumulate successes and failures, to give a better picture of the performance of a system or a human.

6.1. The Interpretability (or Lack Thereof) of Transformers

The success at demonstrating apparent understanding of GPT-3 and its subsequent variations of sizes and styles of transformer networks beg the question of its interpretability and explainability. Consequently, there is emerging work seeking to interpret the internal representations underlying transformers success; it is an active area in which researchers are starting to probe AI understanding and might further benefit from organizing the investigations by the systematic areas of probing outlined above.

Self-attention (Vaswani et al., 2017), the driving force behind the power of the transformer, has come out in front as an interpretable neural network due to its ability to link network weights to specific natural language tokens or pixels in an image; that is, it brings attention to what is important. This view is common in the literature (e.g., Xu et al., 2015; Martins and Astudillo, 2016; Choi et al., 2017; Li et al., 2017; Xie et al., 2017; Vig, 2019; Tang et al., 2020). To quote Li and colleagues: "Attention provides an important way to explain the workings of neural models, at least for tasks with an alignment modeled between inputs and outputs, like machine translation or summarization" (Li et al., 2017, p. 2).

In reality, displaying this interpretability is not as simple as one may be led to believe. However, we posit that attempts to display attention weight relationships for interpretability are an example of attempts to probe the transformer's understanding. Specifically, they are probing the relevant relationships. For example, Jain and Wallace (2019) performed extensive experiments across a variety of NLP tasks that aim to assess the validity of using attention weights as explanations for the network's predictions. They tested two lines of thinking. Attention weights should correlate with feature importance measures, and counterfactual attention weights should lead to corresponding changes in prediction. Their results suggest

that even though these attention models consistently lead to indisputable improved performance on NLP tasks, the transparency, explainability, and interpretability of these models is questionable at best, especially when these models are deep and have complex connections.

Brunner et al. (2020) found similar results in their study of identifiability of attention weights and token embeddings. They found that attention weights are not identifiable, i.e., there are infinitely many attention distributions that can lead to the same internal representation and model output. However, they present Hidden Token Attribution, a gradient-based method to quantify information mixing and showcase its ability to investigate contextual embeddings in self-attention models. It seems hope is not lost on the interpretability of transformers. Chefer et al. (2021) recognize the difficulty in following connections of complex networks and have benchmarked their method on recent visual Transformer networks (such as ViT model), as well as on text classification problems (BERT). They have demonstrated the validity of their approach over existing explainability methods. In the world of transformers and attention, the question of understanding is still up for debate.

7. DISCUSSION AND CONCLUSION

We have outlined herein a set of natural language probe structures that can be adapted to different domains and applied to both human and AI understanding. Critical to evaluating theories about understanding, these can be defined independently of proposed theories and prior to any empirical evaluations. They provide the structure for independent evaluations. They also have the flexibility to adapt to different contexts for assessing understanding to provide a consistently measured body of evidence. Thus, consistent with Hannon (2021)'s recent argument, we can craft that set of criteria to define understanding through the various degrees and abilities (plural) enabled by the process of understanding.

We have argued here that natural language is the core method for probing understanding. We have highlighted that while there are many ways of showing understanding (e.g., performing well on a task), we are suggesting that language, because it is the most familiar symbolic system to humans, is the best, if not the only, method for probing understanding. We should highlight that by natural language we do not mean "perfect spoken language." First, we realize that language can be extremely nuanced with voice tone, gesture, etc. Second, there are many forms of language that can convey many of the same signals—sign language, text, etc. Forms of language that can take advantage of multi-modal cues may convey understanding with more efficient communications. Thus, we are proposing that the more language-cues (e.g., spontaneous gesture, intonation) that are available, the more nuanced and better probes of understanding will be.

Additional complexity in structuring probes for elucidating understanding arises because sometimes we are probing understanding of the external world or mechanical systems, and sometimes we are probing an agent's understanding of another human or intelligent system, as well as whether teamed intelligent agents share mutual understanding. The process of understanding has a flexibility that can support reasoning and successful interpretation of all these types. Probes will need to flexibly adapt, because probes designed for one type of understanding may not elucidate another. In the present work, we have not yet outlined a way to translate the probe structures into specific experimental paradigms. There is likely not a single way to do this; it will depend on a number of factors, like whether you are probing humans or intelligent agents, whether you have spoken or strictly typed communications (or a combination of modalities), and whether the probes are only posed in conversation/communication tasks or if there are additional task completion targets or performance metrics to pair with the probes. Elaborating potential paradigms for putting the probes into practice is left for future work.

There remain some intelligent behaviors that systematic probing may still struggle to help measure or explain as the process of understanding unfolds. Consider the sudden ability to solve an insight problem (Metcalfe, 1986; Metcalfe and Wiebe, 1987). People are generally unable to articulate how they are trying to reason through or solve a problem prior to insight. After the "aha" moment however, people can explain the solution verbally. This is further evidence that understanding requires natural language expression. Not enough of the process has unfolded when the person cannot explain their understanding; the ability to articulate understanding marks achieving a depth of understanding that can be probed.

One possible critique of our proposal that natural language is the core method for probing understanding is that understanding can be demonstrated by performance. For example, if a robot observes a tennis player and learns how to hit various tennis shots, does it understand how to play tennis? In this scenario, the robot could have simply learned various cues for how to hit the ball (stimulus-response) or even how to move itself to win a point. However, we would argue that unless it could use symbolic communication—language of some sort—it does not actually understand the game of tennis (or even the shots it can make). For example, if the robot could describe why it would lob a ball over a net player, we would judge it to have a much better understanding of the game than if the robot could just perform the action at the right time. Along these lines, Baker et al. (2020) demonstrated the emergence of intelligent behaviors in reinforcement learning agents that did not have any NLP capabilities. This seems to be a counter argument to our natural language requirement. While the agents do move through several levels of sophistication in their coordinated activities, they do this with perfect internal knowledge of the states of each other and the environment. Take away any of this knowledge, and the coordination will falter. This suggests that the need to establish understanding within and between agents is the consequence of humans and most systems lacking perfect knowledge of the states of the other agents. That information must be communicated through a common symbolic system.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LB coordinated the paper and wrote Section 1, introduction and attempts at understanding definitions. SK contributed to the introduction and overall argument structure. MA and CB wrote the Sections 2 and 3 on natural language processing. SB and JT wrote Section 4 on common ground and perceived understanding. BH and JK wrote Section 5 on mental models and theory of mind. RS wrote Section 6 on XAI, with CH contributing Section 6.1 on transformers. RW contributed to the discussion, Section 7. All authors contributed to the development of the main hypotheses and to reviewing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was sponsored by the U.S. Department of Defense.

ACKNOWLEDGMENTS

The authors thank Patrick Dull and Austin Blodgett for discussions of the ideas developed in this paper. They thank Glenn Gunzelmann and two reviewers for their helpful comments and suggestions. Distribution A: Cleared for Public Release AFRL-2021-3820.

REFERENCES

Admoni, H., Hayes, B., Feil-Seifer, D., Ullman, D., and Scassellati, B. (2013). "Are you looking at me? Perception of robot attention is mediated by gaze type and group size," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Tokyo, Japan: IEEE), 389–395. doi: 10.1109/HRI.2013.6483614

Allen, J. F., and Perrault, C. R. (1980). Analyzing intention in utterances. *Artif. Intell.* 15, 143–178. doi: 10.1016/0004-3702(80)90042-9

Arimoto, T., Yoshikawa, Y., and Ishiguro, H. (2014). "Nodding responses by collective proxy robots for enhancing social telepresence," in *Proceedings of the Second International Conference on Human-Agent Interaction* (Tsukuba), 97–102. doi: 10.1145/2658861.2658888

Arp, R., Smith, B., and Spear, A. D. (2015). *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262527811.001.0001

Austin, J. L. (1962). *How To Do Things With Words*, Vol. 88. Oxford, UK: Oxford University Press.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2020). "Emergent tool use from multi-agent autocurricula," in *Proceedings of International Conference on Learning Representations (ICLR) 2020* (Virtual), arXiv:1909.07528.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., et al. (2013). "Abstract meaning representation for sembanking," in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (Sofia), 178–186.

Bangalore, S., Hakkani-Tür, D., and Tur, G. (2006). Introduction to the special issue on spoken language understanding in conversational systems. *Speech Commun.* 3, 233–238. doi: 10.1016/j.specom.2005.09.001

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Beaudoin, C., Leblanc, E. L., Gagner, C., and Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Front. Psychol.* 10, 2905. doi: 10.3389/fpsyg.2019.02905

Benninghoff, B., Kulms, P., Hoffmann, L., and Kramer, N. C. (2013). Theory of mind in human-robot-communication: appreciated or not? *Kognitive Systeme*. doi: 10.17185/duepublico/31357

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*. New York, NY: McKay.

Bobrow, D. G., and Collins, A. (eds.). (1975). *Representation and Understanding: Studies in Cognitive Science*. New York, NY: Academic Press, Inc.

Bonial, C., Donatelli, L., Abrams, M., Lukin, S. M., Tratz, S., Marge, M., et al. (2020). "Dialogue-AMR: abstract meaning representation for dialogue," in *Proceedings of the 12th Language Resources and Evaluation Conference* (Marseille), 684–695.

Bonial, C. N., Donatelli, L., Ervin, J., and Voss, C. R. (2019). Abstract meaning representation for human-robot dialogue. *Proc. Soc. Comput. Linguist.* 2, 236–246. doi: 10.18653/v1/W19-3322

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Memory Cogn.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Vancouver, BC), 1877–1901. arXiv:2005.14165.

Bruce, B. C. (1975). "Generation as a social action," in *TINLAP '75: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, eds B. L. Nash-Webber and R. Schank (Stroudsburg, PA: Association for Computational Linguistics), 64–67. doi: 10.3115/980190.980213

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2020). "On identifiability in transformers," in *Proceedings of International Conference on Learning Representations (ICLR) 2020*, arXiv:1908.04211.

Bryant, L., Coffey, A., Povinelli, D. J., and Pruett, J.ohn R., J. (2013). Theory of mind experience sampling in typical adults. *Conscious. Cogn.* 22, 697–707. doi: 10.1016/j.concog.2013.04.005

Buchanan, B. G., and Shortliffe, E. H. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., et al. (2012). "ISO 24617-2: a semantically-based standard for dialogue annotation," in *LREC* (Istanbul: Citeseer), 430–437.

Byom, L. J., and Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Front. Hum. Neurosci.* 7, 413. doi: 10.3389/fnhum.2013.00413

Cannon-Bowers, J., Salas, E., and Converse, S. (1993). "Shared mental models in expert team decision making," in *Individual and Group Decision Making: Current Issues*, ed N. J. Castellan (Hillsdale, NJ: Lawrence Erlbaum Associates), 221–242.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News* 538, 20–23. doi: 10.1038/538020a

Chai, J. Y., Fang, R., Liu, C., and She, L. (2017). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine* 37, 32–45. doi: 10.1609/aimag.v37i4.2684

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., et al. (2014). "Collaborative effort towards common ground in situated human-robot dialogue," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld: ACM), 33–40. doi: 10.1145/2559636.2559677

Chakraborti, T., Sreedharan, S., and Kambhampati, S. (2020). The emerging landscape of explainable AI planning and decision making. *arXiv preprint arXiv:2002.11697*. doi: 10.24963/ijcai.2020/669

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*. doi: 10.1109/CVPR46437.2021.00084

Chen, Y.-N., Hakkani-Tür, D., Tür, G., Gao, J., and Deng, L. (2016). "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding," in *Interspeech 2016* (San Francisco, CA), 3245–3249. doi: 10.21437/Interspeech.2016-312

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J. (2017). "Retain: an interpretable predictive model for healthcare using reverse time attention mechanism," in *29th Conference on Neural Information Processing Systems (NIPS 2016)* (Barcelona). doi: 10.5555/3157382.3157490

Chomsky, N. (1980). Rules and representations. *Behav. Brain Sci*. 3, 1–15. doi: 10.1017/S0140525X00001515

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Clark, H. H. (1994). Managing problems in speaking. *Speech Commun*. 15, 243–250. doi: 10.1016/0167-6393(94)90075-2

Clark, H. H., and Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on Socially Shared Cognition*, eds L. B. Resnick, J. M. Levine, and S. D. Teasley (Hyattsville, MD: American Psychological Association), 127–149. doi: 10.1037/10096-006

Clark, H. H., and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Lang. Cogn. Process*. 2, 19–41. doi: 10.1080/01690968708406350

Clark, H. H., and Schaefer, E. F. (1989). Contributing to discourse. *Cogn. Sci*. 13, 259–294. doi: 10.1207/s15516709cog1302_7

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Cohen, P. R., and Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cogn. Sci*. 3, 177–212. doi: 10.1207/s15516709cog0303_1

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery* 11, e1391. doi: 10.1002/widm.1391

Davis, E. (2011). How does a box work? A study in the qualitative dynamics of solid objects. *Artif. Intell*. 175, 299–345. doi: 10.1016/j.artint.2010.04.006

De Saussure, F. (2011). *Course in General Linguistics*. New York, NY: Columbia University Press.

de Weerd, H., Verbrugge, R., and Verheij, B. (2017). Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Auton. Agents Multiagent Syst*. 31, 250–287. doi: 10.1007/s10458-015-9317-1

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., et al. (2021). Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev*. 54, 755–810. doi: 10.1007/s10462-020-09866-x

Duffield, C. J., Hwang, J. D., Brown, S. W., Dligach, D., Vieweg, S., Davis, J., et al. (2007). "Criteria for the manual grouping of verb senses," in *Proceedings of the Linguistic Annotation Workshop* (Prague), 49–52. doi: 10.3115/1642059.1642067

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). "Retrofitting word vectors to semantic lexicons," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO: Association for Computational Linguistics), 1606–1615. doi: 10.3115/v1/N15-1184

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7287.001.0001

Fillmore, C. J. (1988). "The mechanisms of "construction grammar"," in *Annual Meeting of the Berkeley Linguistics Society* (Berkeley, CA), 35–55. doi: 10.3765/bls.v14i0.1794

Fillmore, C. J., Baker, C. F., and Sato, H. (2002). "The framenet database and software tools," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (Las Palmas)

Fillmore, C. J., Lee-Goldman, R., and Rhodes, R. (2012). "The framenet constructicon," in *Sign-Based Construction Grammar*, eds H. C. Boas and I. A. Sag (Stanford, CA: CSLI), 309–372.

Fox, M., Long, D., and Magazzeni, D. (2017). "Explainable planning," in *Proceedings of IJCAI-17 Workshop on Explainable AI* (Melbourne, VIC), arXiv:1709.10256.

Gentner, D., and Stevens, A. L. (2014). *Mental Models*. New York, NY: Psychology Press. doi: 10.4324/9781315802725

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press.

Gonsior, B., Wollherr, D., and Buss, M. (2010). "Towards a dialog strategy for handling miscommunication in human-robot dialog," in *19th International Symposium in Robot and Human Interactive Communication* (IEEE), 264–269. doi: 10.1109/ROMAN.2010.5598618

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Commun. ACM* 61, 56–66. doi: 10.1145/3134599

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics 3: Speech Acts*, eds P. Cole and J. L. Morgan (London: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Gwern (2020). *Gpt-3 Creative Fiction*. Retrieved from: https://www.gwern.net/GPT-3

Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L., et al. (2016). "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *Interspeech* (San Francisco, CA), 715–719. doi: 10.21437/Interspeech.2016-402

Hannon, M. (2021). Recent work in the epistemology of understanding. *Am. Philos. Q*. 58, 269–290. doi: 10.2307/48616060

Hearst, M. A. (1995). "Tilebars: visualization of term distribution information in full text information access," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM Press/Addison-Wesley Publishing Co.), 59–66. doi: 10.1145/223904.223912

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., and Michalek, H. (2006). *General Formal Ontology (GFO)–A Foundational Ontology Integrating Objects and Processes [Version 1.0*] (Leipzig).

Hirst, G., McRoy, S., Heeman, P., Edmonds, P., and Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Commun*. 15, 213–229. doi: 10.1016/0167-6393(94)90073-6

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hough, A. R., and Gluck, K. A. (2019). The understanding problem in cognitive science. *Adv. Cogn. Syst*. 8, 13–32. Available online at: http://www.cogsys.org/journal/volume8/article-8-3.pdf

Huang, C.-M., and Mutlu, B. (2013). "Modeling and evaluating narrative gestures for humanlike robots," in *Proceedings of the Robotics: Science and Systems Conference (RSS2013)* (Berlin), 57–64. doi: 10.15607/RSS.2013.IX.026

Issar, S., and Ward, W. (1993). "CMU's robust spoken language understanding system," in *Third European Conference on Speech Communication and Technology* (Lisbon).

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Jain, S., and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

James, W. (1890). *The Principles of Psychology*, Vol. 1. New York, NY: Henry Holt & Co. doi: 10.1037/10538-000

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Jones, S. S. (2009). The development of imitation in infancy. *Philos. Trans. R. Soc. B Biol. Sci*. 364, 2325–2335. doi: 10.1098/rstb.2009.0045

Jonker, C. M., Van Riemsdijk, M. B., and Vermeulen, B. (2010). "Shared mental models," in *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems* (Berlin, Heidelberg: Springer), 132–151. doi: 10.1007/978-3-642-21268-0_8

Jurafsky, D., and Martin, J. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kennedy, W. G., Bugajska, M. D., Adams, W., Schultz, A. C., and Trafton, J. G. (2008). "Incorporating mental simulation for a more effective robotic teammate," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 1300–1305.

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 23, 89–109. doi: 10.1016/S0933-3657(01)00077-X

Korpan, R., and Epstein, S. L. (2018). "Toward natural explanations for a robot's navigation plans," in *Notes from the Explainable Robotic Systems Worshop, Human-Robot Interaction 2018*, eds M. de Graaf, B. Malle, A. Dragan, and T. Ziemke (Chicago, IL).

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: an overview. *Theory Into Practice* 41, 212–218. doi: 10.1207/s15430421tip4104_2

Leslie, A. M., Knobe, J., and Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychol. Sci.* 17, 421–427. doi: 10.1111/j.1467-9280.2006.01722.x

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition* 14, 41–104. doi: 10.1016/0010-0277(83)90026-4

Li, J., Monroe, W., and Jurafsky, D. (2017). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Martins, A. F. T., and Astudillo, R. F. (2016). "From softmax to sparsemax: a sparse model of attention and multi-label classification," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (New York, NY), 1614–1623.

Matuszek, C., Witbrock, M., Cabral, J., and DeOliveira, J. (2006). "An introduction to the syntax and content of Cyc," in *Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, Papers from the 2006 AAAI Spring Symposium* (Menlo Park, CA: AAAI Press).

McCarthy, J. (1990). "An example for natural language understanding and the AI problems it raises," in *Formalizing Common Sense: Papers by John McCarthy*, ed V. Lifschitz (Norwood, NJ. Ablex Publishing Corporation), 70–76.

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM Sigart Bull.* 57, 4–9. doi: 10.1145/1045339.1045340

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Dev. Psychol.* 31, 838–850. doi: 10.1037/0012-1649.31.5.838

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12, 623–634.

Metcalfe, J. and Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition* 15, 238–246.

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Memory Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Michaelis, L. A., and Lambrecht, K. (1996). Toward a construction-based theory of language function: the case of nominal extraposition. *Language* 72, 215–247. doi: 10.2307/416650

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Mills, G. J. (2014). Dialogue in joint activity: complementarity, convergence and conventionalization. *N. Ideas Psychol.* 32, 158–173. doi: 10.1016/j.newideapsych.2013.03.006

Moore, J., and Newell, A. (1974). "How can merlin understand?" in *Cognition and Knowledge*, ed L. W. Gregg (Potomac, MD: Lawrence Erlbaum Associates), 201–252.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

Mumm, J., and Mutlu, B. (2011). "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne), 331–338. doi: 10.1145/1957656.1957786

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., and Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.* 1, 1–33. doi: 10.1145/2070719.2070725

Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. (2019). "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Menlo Park, CA: AAAI Press), 97–105.

O'Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K., and Palmer, M. (2018). "AMR beyond the sentence: the multi-sentence AMR corpus," in *Proceedings of the 27th International Conference on Computational Linguistics* (Santa Fe), 3693–3702.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds Mach.* 29, 441–459. doi: 10.1007/s11023-019-09502-w

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: an annotated corpus of semantic roles. *Comput. Linguist.* 31, 71–106. doi: 10.1162/0891201053630264

Pomerleau, D. A. (1992). "Progress in neural network-based vision for autonomous robot driving," in *Proceedings of the Intelligent Vehicles Symposium* (New York, NY: IEEE), 391–396.

Potts, C. (2012). "Goal-driven answers in the cards dialogue corpus," in *Proceedings of the 30th West Coast Conference on Formal Linguistics* (Somerville, MA), 1–20.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). "Ontonotes: a unified relational semantic representation," in *International Conference on Semantic Computing (ICSC 2007)*. (Irvine, California, USA: IEEE), 517–526. doi: 10.1109/ICSC.2007.83

Rouse, W. B., and Morris, N. M. (1986). On looking into the black box: prospects and limits in the search for mental models. *Psychol. Bull.* 100, 349–363. doi: 10.1037/0033-2909.100.3.349

Salas, E., Stout, R., and Cannon-Bowers, J. (1994). "The role of shared mental models in developing shared situational awareness," in *Situational Awareness in Complex Systems*, eds R. D. Gilson, D. J. Garland, and J. M. Koonce (Daytona Beach, FL: Embry-Riddle Aeronautical University Press), 297–304.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382. doi: 10.1353/lan.1977.0041

Schubert, L. K. (2015). "Semantic representation," in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, TX), 4132–4138.

Scielzo, S., Fiore, S. M., Cuevas, H. M., and Salas, E. (2004). "Diagnosticity of mental models in cognitive and metacognitive processes: Implications for synthetic task environment training," in *Scaled Worlds: Development, Validation, and Applications*, eds L. R. Elliott and M. D. Coovert (Aldershot: Ashgate), 181–199.

Searle, J. (1984). "Can computers think?" in *Minds, Brains, and Science*, ed J. Searle (Cambridge, MA. Harvard University Press), 28–41.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*, Vol. 626. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781139173438

Shieber, S. M. (1994). Lessons from a restricted turing test. *arXiv preprint arXiv: cmp-lg/9404002*. doi: 10.1145/175208.175217

Sidner, C. L., Lee, C., Morency, L.-P., and Forlines, C. (2006). "The effect of head-nod recognition in human-robot conversation," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City), 290–296. doi: 10.1145/1121241.1121291

Simon, H. A. (1977). "Artificial intelligence systems that understand," in *IJCAI* (Cambridge, MA), 1059–1073.

Simon, H. A., and Eisenstadt, S. A. (2000). *A Chinese Room that Understands*. Pittsburgh, PA: Carnegie Mellon University.

Simon, H. A., and Hayes, J. R. (1976). The understanding process: problem isomorphs. *Cogn. Psychol.* 8, 165–190. doi: 10.1016/0010-0285(76)90022-0

Stalnaker, R. (2002). Common ground. *Linguist. Philos.* 25, 701–721. doi: 10.1023/A:1020867916902

St. Amant, R., Fields, M., Kaukeinen, B., and Robison, C. (2019). "Lightweight schematic explanations of robot navigation," in *Proceedings of the International Conference on Cognitive Modeling (ICCM)* (Montreal, QC).

Steedman, M., and Baldridge, J. (2011). "Combinatory categorial grammar," in *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, eds R. D. Borsley and K. Börjars (Oxford: Blackwell), 181–224. doi: 10.1002/9781444395037.ch5

Summers-Stay, D., Bonial, C., and Voss, C. (2021). "What can a generative language model answer about a passage?" in *The 3rd Workshop on Machine Reading for Question Answering* (Punta Cana), doi: 10.18653/v1/2021.mrqa-1.7

Takayama, L., and Pantofaru, C. (2009). "Influences on proxemic behaviors in human-robot interaction," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (St Louis, MO: IEEE), 5495–5502. doi: 10.1109/IROS.2009.5354145

Tang, Y., Nguyen, D., and Ha, D. (2020). "Neuroevolution of self-interpretable agents," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference* (New York, NY). doi: 10.1145/3377930.3389847

Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., and Nivel, E. (2016). "About understanding," in *International Conference on Artificial General Intelligence* (Cham: Springer), 106–117. doi: 10.1007/978-3-319-41649-6_11

Trafton, J. G., Bugajska, M. D., Fransen, B. R., and Ratwani, R. M. (2008). "Integrating vision and audition within a cognitive architecture to track conversations," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 201–208. doi: 10.1145/1349822.1349849

Traum, D. R. (1999). "Speech acts for dialogue agents," in *Foundations of Rational Agency*, eds A. Rao and M. Wooldridge (Kluwer) (Dordrecht: Springer), 169–201. doi: 10.1007/978-94-015-9204-8_8

Turing, A. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach CA: ACM), 6000–6010. doi: 10.5555/3295222.3295349

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*. doi: 10.18653/v1/P19-3007

Vilone, G., and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Vossen, P. (1997). "Eurowordnet: a multilingual database for information retrieval," in *Proceedings of the DELOS Workshop on Cross-Language Information Retrieval* (Zurich: Vrije Universiteit).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). "GLUE: a multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Brussels), 353–355. doi: 10.18653/v1/W18-5446

Weigand, E. (1999). Misunderstanding: the standard case. *J. Pragmat*. 31, 763–785. doi: 10.1016/S0378-2166(98)00068-X

Weld, D. S., and Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 70–79. doi: 10.1145/3282486

Wilpon, J. G., Rabiner, L. R., Lee, C.-H., and Goldman, E. (1990). Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Trans. Acoust. Speech Signal Process.* 38, 1870–1878. doi: 10.1109/29.103088

Winfield, A. F. T. (2018). Experiments in artificial theory of mind: from safety to story-telling. *Front. Robot. AI* 5, 75. doi: 10.3389/frobt.2018.00075

Woods, W. A. (1973). Progress in natural language understanding: an application to lunar geology," in *Proceedings of the National Computer Conference and Exposition* (New York, NY), 441–450. doi: 10.1145/1499586.1499695

Xie, Q., Ma, X., Dai, Z., and Hovy, E. (2017). An interpretable knowledge transfer model for knowledge base completion. *arXiv preprint arXiv:1704.05908*. doi: 10.18653/v1/P17-1088

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges," in *CCF International Conference on Natural Language Processing and Chinese Computing* (Cham: Springer), 563–574. doi: 10.1007/978-3-030-32236-6_51

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention." in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)* (Lille), 2048–2057. Available online at: https://arxiv.org/abs/1502.03044

Young, R. M., Pollack, M. E., and Moore, J. D. (1994). "Decomposition and causality in partial-order planning," in *International Conference on Artificial Intelligence Planning Systems* (Menlo Park, CA: AAAI Press), 188–194.

Zhong, Z., Ng, H. T., and Chan, Y. S. (2008). "Word sense disambiguation using ontonotes: an empirical study," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI), 1002–1010. doi: 10.3115/1613715.1613845

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 12 May 2022
doi: 10.3389/fnsys.2022.845177

Multisensory Concept Learning Framework Based on Spiking Neural Networks

Yuwei Wang[1,2] *and Yi Zeng*[1,2,3,4]*

[1] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, [4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Concept learning highly depends on multisensory integration. In this study, we propose a multisensory concept learning framework based on brain-inspired spiking neural networks to create integrated vectors relying on the concept's perceptual strength of auditory, gustatory, haptic, olfactory, and visual. With different assumptions, two paradigms: Independent Merge (IM) and Associate Merge (AM) are designed in the framework. For testing, we employed eight distinct neural models and three multisensory representation datasets. The experiments show that integrated vectors are closer to human beings than the non-integrated ones. Furthermore, we systematically analyze the similarities and differences between IM and AM paradigms and validate the generality of our framework.

Keywords: concept learning, multisensory, spiking neural networks, brain-inspired, Independent Merge, Associate Merge

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Laxmi R. Iyer,
*Institute for Infocomm Research (A*STAR), Singapore*
Sun Zhe,
RIKEN, Japan
Shangbin Chen,
Huazhong University of Science and Technology, China

***Correspondence:**
Yi Zeng
yi.zeng@ia.ac.cn

Received: *29 December 2021*
Accepted: *20 April 2022*
Published: *12 May 2022*

Citation:
Wang Y and Zeng Y (2022) Multisensory Concept Learning Framework Based on Spiking Neural Networks. Front. Syst. Neurosci. 16:845177. doi: 10.3389/fnsys.2022.845177

1. INTRODUCTION

Concept learning, or the ability to recognize commonalities and accentuate contrasts across a group of linked events in order to generate structured knowledge, is a crucial component of cognition (Roshan et al., 2001). Multisensory integration benefits concept learning (Shams and Seitz, 2008) and plays an important role in semantic processing (Xu et al., 2017; Wang et al., 2020). For example, when we learn the concept of "tea," acoustically, we will perceive the sound of pouring water and brewing, the sound of clashing porcelain, the sound of drinking tea; on taste, we can feel the tea is a bit bitter, astringent or sweet; in touch, tea is liquid and we can feel its temperature; on smell, we can perceive the faint scent and visually, it often appears together with the teapot or tea bowl, and the tea leaves will have different colors. Combining information from multiple senses can produce enhanced perception and learning, faster response times, and improved detection, discrimination, and recognition capabilities (Calvert and Thesen, 2004). In the brain, multisensory integration occurs mostly in the superior colliculus according to existing studies (Calvert and Thesen, 2004; Cappe et al., 2009). Multisensory integration is a field that has attracted the interest of cognitive psychologists, biologists, computational neuroscientists, and artificial intelligence researchers. The term "multisensory concept learning" is used in this work to describe the process of learning concepts using a model that mimics humans and combines information from multiple senses.

For the computational models of multisensory integration, cognitive psychologists' models are usually focused on model design and validation from the mechanism of multisensory integration. These models are highly interpretable, taking neuroimaging and behavioral studies

into consideration. The cue combination model based on Bayesian decision theory is a classical model for analyzing multisensory integration in cognitive psychology. It mainly models the stimuli of different modalities as the likelihood functions of Gaussian (Ursino et al., 2009, 2014) or Poisson (Anastasio et al., 2014) distributions with different parameters, and calculates the best combination of each modality that makes the maximum posterior distribution through the assumption of conditional independence and Bayesian rules. Anastasio et al. built a model of visual and auditory fusion that combines neuronal dynamic equations with feedback information, and this model verified that multimodal stimuli have less response time than unimodal stimuli (Anastasio et al., 2014). Parise et al. proposed multisensory correlation detector based models to describe correlation, lag, and synchrony across the senses (Parise and Ernst, 2016). A purely visual haptic prediction model is presented by Gao et al. (2016) with CNNs and LSTMs, which enables robots to "feel" without physical interaction. Gepner et al. (2015) developed a linear-nonlinear-Poisson cascade model that incorporates information from olfaction and vision to mimic Drosophila larvae navigation decisions, and the model was able to predict Drosophila larvae reaction to new stimulus patterns well.

For artificial intelligence researchers, they have proposed different types of multisensory integration models based on the available data and machine learning methods, such as direct concatenation (Kiela and Bottou, 2014; Collell et al., 2017; Wang et al., 2018b), canonical correlation analysis (Silberer et al., 2013; Hill et al., 2014), singular value decomposition of the integration matrix (Bruni et al., 2014), multisensory context (Hill and Korhonen, 2014), autoencoders (Silberer and Lapata, 2014; Wang et al., 2018a), and tensor fusion networks (Zadeh et al., 2017; Liu et al., 2018; Verma et al., 2019). These works are mostly focused on concept learning and sentiment analysis tasks and are based on modeling of speech, text, and image data, which are commonly utilized in AI.

To our knowledge, no work exists to model the five senses of vision, hearing, touch, taste, and smell together. This might be because controlling elements for experimental design is challenging for cognitive psychologists, while data for some modalities is difficult to get using perceptrons for AI researchers. Meanwhile, cognitive psychologists have published several multisensory datasets by asking volunteers how much they perceive a specific concept through their auditory, gustatory, tactile, olfactory, and visual senses in order to establish the strength of each modality. This provides a solid basis for the design of a multisensory integration model that includes these five modalities. In this article, we will model multisensory integration using brain-like spiking neural networks and merge input from five different modalities to generate integrated representations.

This paper is organized as follows: Section 2 will introduce relevant studies to our model, such as multisensory datasets and fundamental SNN models; Section 3 will describe the multisensory concept learning framework based on SNNs, which includes the Independent Merge and Associate Merge paradigms. Section 4 will exhibit the experiments, and the final section will explore the future works.

2. RELATED WORKS

2.1. Multisensory Concept Representation Datasets

Cognitive psychologists label the multisensory datasets of concepts by asking volunteers how much each concept is acquired through a specific modality and introducing statistical methods to establish the representation vector for each concept. The pioneering work in this area is by Lynott and Connell (2013), who proposed modality exclusivity norms for 423 adjective concepts (Lynott and Connell, 2009) and 400 nominal concepts on strength of association with each of the five primary sensory modalities (auditory, gustatory, haptic, olfactory, visual). In this article, we combine these two datasets of their previous works and denote them as LC823. Lancaster Sensorimotor Norms were published by Lynott et al. (2019), which included six perceptual modalities (auditory, gustatory, haptic, interoceptive, olfactory, visual) and five action effectors (foot/leg, hand/arm, head, mouth, torso). This dataset (we denote as Lancaster40k) is the largest ever, with 39,707 psycholinguistic concepts (Lynott et al., 2019). Binder et al. (2016) constructed a set of brain-based componential semantic representation (BBSR) with 65 experienced attributes, including sensory, motor, spatial, temporal, affective, social, and cognitive experiences, relying on more recent neurobiological findings. This dataset contains 535 concepts and does an excellent work of separating a priori conceptual categories and capturing semantic similarity (Binder et al., 2016). **Figure 1** shows the the concept "honey" in the multisensory concept representation datasets mentioned.

We'll concentrate on the effect of five forms of senses in this article: vision, touch, sound, smell, and taste. In BBSR, we employ the average value of the sub-dimensions corresponding to these five senses, while using the first five dimensions of Lancaster40k.

2.2. Basic Neuron and Synapse Models

Spiking neural networks (SNNs) are commonly referred to be the third generation of neural network models since theyăareăinspired by current discoveries in neuroscience (Maass, 1997). Neurons are the basic processing units of the brain. They communicate with each other *via* synapses. When the membrane potential reaches a threshold, a spike is produced. External stimuli are conveyed by firing rate and the temporal pattern of spike trains (Rieke et al., 1999; Gerstner and Kistler, 2002). SNNs integrate temporal information into the model and are capable of accurately describing spike timing with dynamic changes in synaptic weights which are more biologically plausible. We will use SNNs as the foundation of our model to build a human-like multisensory integration concept learning framework. Here, we briefly outline the neural and synaptic models that will be used in this research.

2.2.1. IF Neural Model

The integrate-and-fire (IF) model is a large family of models which assumes that a membrane potential threshold controls the

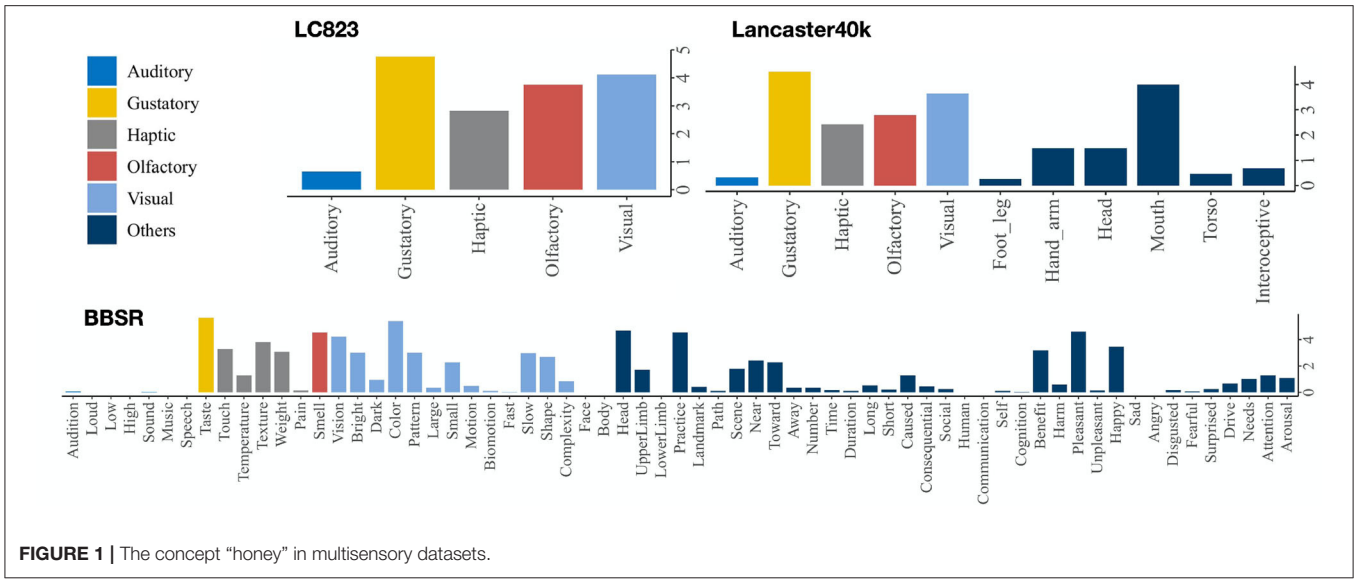


FIGURE 1 | The concept "honey" in multisensory datasets.

spikes of neurons. A spike is fired when the somatic membrane potential exceeds the threshold, and the membrane potential is resumed to reset potential (Gerstner and Kistler, 2002). The neural processing is properly formalized by the model. In this article, we follow a standard implementation (Troyer and Miller, 1997), and the membrane potential $v(t)$ obeys

$$\begin{aligned} \tau_{IF}\frac{dv(t)}{dt} &= v_{rest} - v(t) + g_e(t)(E_e - v(t)) \\ if\ \ v(t) &> v_{th},\ \ v(t) \leftarrow v_r \end{aligned} \tag{1}$$

with the membrane time constant $\tau_{IF} = 20\ ms$, the resting potential $v_{rest} = -14\ mV$, the threshold for spike firing $v_{th} = 6\ mV$, the reset potential $v_r = 0\ mV$, and excitatory potential $E_e = 0\ mV$. Synaptic inputs are modeled as conductance g_e changes with $\tau_e \frac{dg_e}{dt} = -g_e$, where $\tau_e = 5\ mV$.

2.2.2. LIF Neural Model

The leaky integrate-and-fire (LIF) neuron model is one of the most popular spiking neuron models because it is biologically realistic and computationally easy to study and mimic. The LIF neuron's subthreshold dynamics are described by the equation below:

$$\begin{aligned} \tau_{LIF}\frac{dv(t)}{dt} &= v_{rest} - v(t) + R_m I \\ if\ v(t) &> v_{th},\ v(t) \leftarrow v_r \end{aligned} \tag{2}$$

In this paper, the membrane resistance constance $R_m = 1$, $\tau_{LIF} = 20$, $v_{rest} = 1.05$, $v_{th} = 1$, and $v_r = 0$.

2.2.3. Izhikevich Neural Model

Izhikevich model was first proposed in 2003 to replicate spiking and bursting behavior of known types of cortical neurons. The model combines the biological plausibility of Hodgkin and Huxley (1952) dynamics with the computing efficiency of integrate-and-fire neurons (Izhikevich, 2003). Biophysically accurate Hodgkin-Huxley neural models are reduced to a two-dimensional system of the following dynamics ordinary with bifurcation methods:

$$\begin{aligned} \frac{dv(t)}{dt} &= 0.04v(t)^2 + 5v(t) + 140 - u(t) + I, \\ \frac{du}{dt} &= a(bv(t) - u(t)) \\ if\ v(t) &> v_{th},\ v(t) \leftarrow c\ \ and\ \ u(t) \leftarrow u(t) + d \end{aligned} \tag{3}$$

where the time scale of the recovery variable u is described by the parameter a, the sensitivity of the recovery variable u to subthreshold changes of the membrane potential v is described by the parameter b, the parameter c defines the membrane potential v's after-spike reset value, which is induced by quick high-threshold K^+ conductances and after-spike reset of the recovery variable u induced by slow high-threshold Na^+ and K^+ conductances is described by the parameter d (Izhikevich, 2003).

The model simulates the spiking and bursting activity of known kinds of cortical or thalamic neurons such as resonator (RZ), fast spiking (FS), intrinsically bursting (IB), low-threshold spiking (LTS), regular spiking (RS), chattering (CH), and thalamo-cortical (TC) based on these four parameters. These

TABLE 1 | Izhikevich models.

Neurons	Izhikevich parameters			
	a	b	c	d
RZ (resonator)	0.10	0.25	−65	2
FS (fast spiking)	0.10	0.20	−65	2
IB (intrinsically bursting)	0.02	0.20	−55	4
LTS (low-threshold spiking)	0.02	0.25	−65	2
RS (regular spiking)	0.02	0.20	−65	8
CH (chattering)	0.02	0.20	−50	2
TC (thalamo-cortical)	0.02	0.25	−65	0.05

models are employed extensively in our work and details are illustrated in **Table 1**.

2.2.4. STDP Synapse Models

Spike-timing-dependent plasticity (STDP) is a biological process that modifies the strength of neural connections in the brain. Learning and information storage in the brain, as well as the growth and refinement of neural circuits throughout brain development, are thought to be influenced by STDP (Bi and Poo, 2001). The typical STDP model is used in this research, and the weight change Δw of a synapse relies on the relative time of presynaptic spike arrivals and postsynaptic spike arrivals. $\Delta w = \Sigma_{t_{pre}} \Sigma_{t_{post}} W(t_{post} - t_{pre})$, where the function $W(\cdot)$ is defined as:

$$W(\Delta t) = \begin{cases} A_+ exp(\frac{\Delta t}{\tau_+}) \ \Delta t > 0 \\ -A_- exp(-\frac{\Delta t}{\tau_-}) \ \Delta t < 0 \end{cases} \tag{4}$$

When implement STDP, we follow the way of Brian2 (Stimberg et al., 2019), which defines two variables a_{pre} and a_{post} as the "traces" of of pre- and post-synaptic activity, governed by the following differential equations

$$\begin{aligned} \tau_{pre} \frac{a_{pre}}{dt} &= -a_{pre} \\ \tau_{post} \frac{a_{post}}{dt} &= -a_{post} \end{aligned} \tag{5}$$

Once a presynaptic spike occurs, the presynaptic trace is updated and the weight is modified according to the rule

$$\begin{aligned} a_{pre} &\leftarrow a_{pre} + A_{pre} \\ w &\leftarrow w + a_{post} \end{aligned} \tag{6}$$

And when a postsynaptic spike occurs:

$$\begin{aligned} a_{post} &\leftarrow a_{post} + A_{post} \\ w &\leftarrow w + a_{pre} \end{aligned} \tag{7}$$

This is proved to be equivalent for the two kinds of STDP formulations. And, in this article $\tau_{pre} = \tau_{post} = 1ms$.

3. THE FRAMEWORK OF MULTISENSORY CONCEPT LEARNING FRAMEWORK BASED ON SPIKING NEURAL NETWORKS

We present a multisensory concept learning framework based on SNNs in this part. The model's input is a multisensory vector labeled by cognitive psychologists, with an integrated vector as the output following SNNs merging. Since there is no biological study to show whether the information of multiple senses is independent or associated before integration, two different paradigms: Independent Merge (IM) and Associate Merge (AM) are designed in our framework. The types of inputs and outputs are the same for both paradigms, but the architectural design of SNNs is different. These two paradigms involve the same phase in the framework, and only oneăparadigm is chosen for concept integration, depending on the assumption that whether multiple sensory input is independent before integration.

Figure 2 illustrates the workflow: Firstly, for each modality of the concept, we employ a neural model and transform its perceptual strength in the concept's multisensory vector into external stimuli to the neuron (we work on five sensory modalities: auditory, gustatory, haptic, olfactory, visual, so the dimensions of the multisensory vector is five); Secondly, the architecture of SNN is designed according to different assumptions. We choose the IM paradigm if we assume that multiple senses are independent of each other before fusion, and we choose the AM paradigm if we assume that multiple senses are associated with each other; Thirdly, we specify the neuron model in SNN and sequentially feed concepts to the network, with STDP rules adjusting the network's connection weights. Given the running interval $[0, T]$, we record the spike trains of each neuron; Finally, we convert the spike trains of specific neurons into binarycode as the final integrated representation. The framework is described in detail with the IM and AM paradigms individually in the following sections.

3.1. The Framework

3.1.1. Independent Merge

The IM paradigm is founded on the commonly used cognitive psychology assumption that information for each modality of the concept is independent before integration. It's a two-layer spiking neural network model, with five neurons corresponding to the stimuli of the concept's five separate modal information in the second layer, and a neuron reflecting the neural state after multisensory integration in the second layer. We record the spiking train of the postsynaptic neuron and transform them into integrated vectors for the concept.

For each concept, we get its representation from human-labeled vectors, $\vec{m} = [m_A, m_G, m_H, m_O, m_V]$. The subscripts here represent the concept's perceptual strength as indicated by auditory, gustatory, haptic, olfactory, and visual senses. We min-max normalize the multisensory representation of the concept in the dataset as input to the model during the data preparation stage such that each value of the vector is between 0 and 1. In LC823, for instance, the vector for the concept "honey" is [0.13, 0.95, 0.57, 0.75, 0.80]. We employ LIF or Izhikevich as presynaptic neural models and IF as postsynaptic neural models independently for the generality of the framework. Initially, for each presynaptic neuron, we regard the current $I = m_i * I_{boost}$ as the stimuli to the neuron *where* $i \in [A, G, H, O, V]$ The the conductance g_e of the postsynaptic neuron is updated whenever the presynaptic neuron fires as $g_e \leftarrow g_e + \Delta W_{ij}$, and the postsynaptic neuron generates spikes based on the IF model. The synaptic strength between the postsynaptic neuron and the presynaptic neuron is referred to as the weight ΔW_{ij} in this case. The initial weights between presynaptic and postsynaptic neurons $W_0^i = \frac{g_i}{\Sigma_i^n g_i}$ where $g_i = \frac{1}{\sigma_i^2}$,$\sigma_i^2$ represents the variance for each kind of multisensory data. They are calculated using the Bayesian formula and the assumption that each modal is independent before to fusion (details in the Appendix). At the same time, the spike trains of presynaptic and postsynaptic neurons will dynamically adjust to the weights in accordance with the STDP law. During $[0, T]$, we record the spike train of the postsynaptic neuron $S^{post}([0, T])$ and transform them into

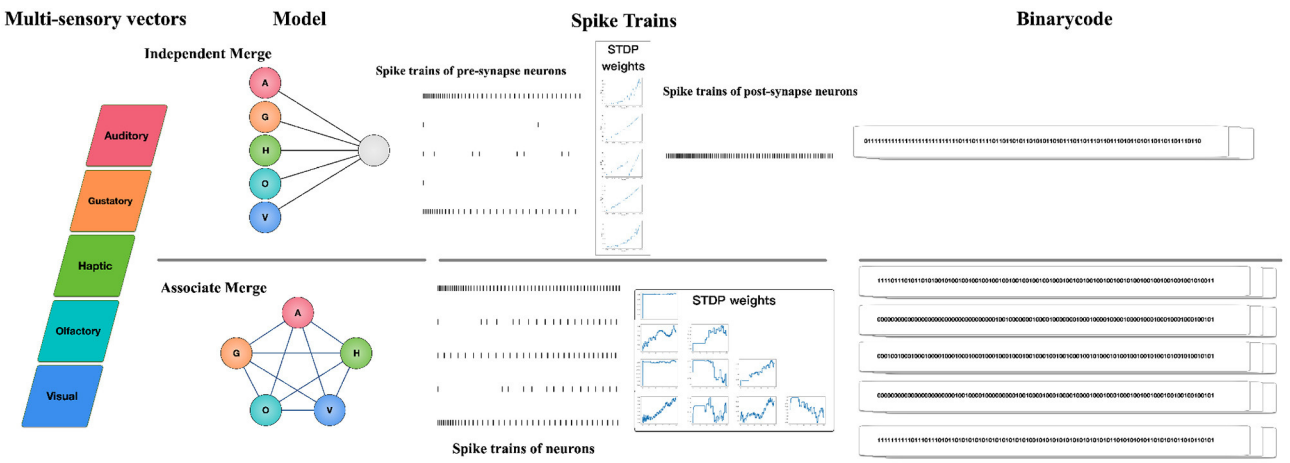


FIGURE 2 | The framework of multisensory concept learning framework based on spiking neural networks. Firstly, for each modality of the concept, we employ a neural model and transform its perceptual strength in the concept's multisensory vector into external stimuli to the neuron; Secondly, the architecture of SNN is designed according to different assumptions. We choose the IM paradigm if we assume that multiple senses are independent of each other before fusion, and we choose the AM paradigm if we assume that multiple senses are associated with each other; Thirdly, we specify the neuron model in SNN and sequentially feed concepts to the network, with STDP rules adjusting the network's connection weights. Given the running interval $[0, T]$, we record the spike trains of each neuron; Finally, we convert the spike trains of specific neurons into binarycode as the final integrated representation.

binarycode $B^{post}([0, T])$, as the final integration representation for the concept in the following manner:

$$B^{post}([0,T]) = [\mathcal{T}(S^{post}((0,tol])), \mathcal{T}(S^{post}((tol, 2*tol])), \cdots, \mathcal{T}(S^{post}(((k-1)*tol, k*tol])), \cdots, \mathcal{T}(S^{post}((\lfloor\frac{T}{tol}\rfloor * tol, T]))] \tag{8}$$

Here $\mathcal{T}(interval)$ operation means that if there is any spikes in the interval, then the bit is 1, otherwise it is 0.

3.1.2. Associate Merge

The AM paradigm assumes that the information for each modality of the concept is associate before integration. It's a five-neuron spiking neural network model, with five neurons corresponding to the stimuli of the concept's five separate modal information. They are connected to one another, and there are no self-connections. We record the spiking trains of all neurons and transform them into integrated vectors for the concept.

We use LIF or Izhikevich neural models to model each neuron for the generality of the framework. For each concept, we get its normalized representation from human-labeled vectors, $\vec{m} = [m_A, m_G, m_H, m_O, m_V]$. Initially, for each neuron $i \in [A, G, H, O, V]$, we consider $I = m_i * I_{boost}$ as the stimuli. The the current I of the postsynaptic neuron is updated whenever the presynaptic neuron fires as $I \leftarrow I + \Delta W_{ij}$. And the postsynaptic neuron generates spikes based on the its model. The weight W_{ij} is the synaptic strength between the presynaptic neuron and the postsynaptic neuron. The initial value for the weight is determined by the correlation each modality pair overall the representation dataset, i.e., $W_0 = Corr(i, j)$ where $i, j \in [A, G, H, O, V]$, which is different from AM paradigm. Simultaneously, presynaptic and postsynaptic neurons' spike trains will dynamically change to the weights in accordance with the STDP law. We denote $S^i([0, T])$ as the *ith* neuron's spike trains during $[0, T]$ and corresponding binary vector $B^i([0, T])$. And we record the spike trains of all neurons, transform them into binarycode $B^i([0, T])$ and concatenate them as the final integration vector $B([0, T])$ in the following way:

$$B^i([0,T]) = [\mathcal{T}(S^i((0,tol])), \mathcal{T}(S^i((tol, 2*tol])), \cdots, \mathcal{T}(S^i(((k-1)*tol, k*tol])), \cdots, \mathcal{T}(S^i((\lfloor\frac{T}{tol}\rfloor * tol, T]))] \tag{9}$$

$$B([0,T]) = [B^A([0,T]) \oplus B^H([0,T]) \oplus B^G([0,T]) \oplus B^O([0,T]) \oplus B^V([0,T])] \tag{10}$$

4. EXPERIMENTS

4.1. Concept Similarity Test

Concept similarity test is commonly used in the field of artificial intelligence to evaluate the effectiveness of system-generated representations (Agirre et al., 2009). Generally, humans score the similarity of a particular concept pair, while the concept pair corresponds to the system-generated representation to calculate the similarity score. After the two scores are ranked in the measure dataset, the Spearman's correlation coefficient is calculated to reflect how close the system-generated representations are to humans. In this article, we evaluate the closeness of the concepts' original or multisensory integration representations and human beings with WordSim353 (Agirre et al., 2009) and SCWS1994 (Huang et al., 2012).

4.1.1. The Experiment

To thoroughly test our framework, we did experiments for IM and AM paradigms with three multisensory datasets

(BBSR, LC823, Lancaster40k) respectively and analyzed the effectiveness differences between the representations after SNN integration and the original representations. In the experiments, both IM and AM paradigms involve a unique parameter in the process of conversion from spike trains to binarycode: the tolerance *tol*. It represents the size of the reducing window for converting spike trains in the time interval into binarycode, which reflects the strength of compressing the spike sequence into a integrated binarycode. In each dimension of the integrated vector, a larger *tol* signifies a higher degree of information compression and a bigger reducing window, and *vice versa*. But, if *tol* is too small, the representation vector's dimensionality will be too large, and if *tol* is too big, the diversity of all representations will be damaged. Therefore, we traverse *tol* across the range $[0, 500]$ while restricting diversity to the range $[0.05, 0.95]$, and the results indicate the present model's ideal results as well as the matching *tol*.

TABLE 2 | Concept similarity test results.

Merge way	Model	BBSR				LC823an				Lancaster40k			
		Tol	WordSim353	SCWS1994	Average	Tol	WordSim353	SCWS1994	Average	Tol	WordSim353	SCWS1994	Average
Origin	–	–	0.4182	0.5838	0.5010	–	0.1321	0.5525	0.3423	–	0.2640	**0.3974**	0.3534
AM	lzh-RZ	93	0.3455	0.6089	0.4772	165	0.3804	0.4260	0.4032	9	0.3560	0.3295	0.3427
	lzh-FS	95	0.4955	0.5659	0.5307	312	0.4223	0.3788	0.4006	9	0.3787	0.3471	0.3629
	lzh-IB	384	0.5455	0.5870	0.5662	32	0.3696	0.5277	0.4486	25	0.3388	0.3818	0.3603
	lzh-LTS	174	0.5068	0.6127	0.5598	17	0.3107	0.5390	0.4249	16	0.3557	0.3629	0.3593
	lzh-RS	366	0.4955	0.5857	0.5406	84	0.5179	0.5271	0.5225	55	0.3206	0.3708	0.3457
	lzh-CH	170	0.4273	0.5928	0.5100	7	0.1089	0.4884	0.2986	14	0.3150	0.3349	0.3249
	lzh-TC	148	0.5068	0.6103	0.5586	6	0.2214	0.5181	0.3698	7	**0.3979**	0.3364	0.3672
	LIF	187	0.5727	0.6927	0.6327	330	0.5036	**0.6330**	0.5683	86	0.1788	0.3500	0.2644
IM	lzh-RZ	17	0.4636	0.634	0.5488	10	0.5545	0.5618	0.5581	4	0.2026	0.3139	0.2583
	lzh-FS	17	0.4636	0.6388	0.5512	10	0.5545	0.5617	0.5581	21	0.3371	0.2910	0.3140
	lzh-IB	7	**0.5477**	0.5988	**0.5733**	24	0.5509	0.5491	0.5500	31	0.1597	0.3040	0.2319
	lzh-LTS	83	0.5000	**0.6417**	0.5708	18	**0.6080**	0.5361	**0.5721**	56	0.3610	0.3448	0.3529
	lzh-RS	196	0.5023	0.5530	0.5276	163	0.4830	0.4613	0.4722	68	0.0757	0.2959	0.1858
	lzh-CH	94	0.4659	0.5786	0.5222	8	0.5696	0.4746	0.5221	50	0.3843	0.3813	**0.3828**
	lzh-TC	17	0.4636	0.6125	0.5381	5	0.4509	0.5310	0.4909	20	0.3387	0.3042	0.3215
	LIF	143	0.4205	0.6167	0.5186	3	0.0643	0.5672	0.3158	324	0.0018	0.1481	0.1965

The bold values indicates the current measure dataset reflect the best results, while the underlined values imply that the multisensory integrated representation is closer to humans than the original representation.

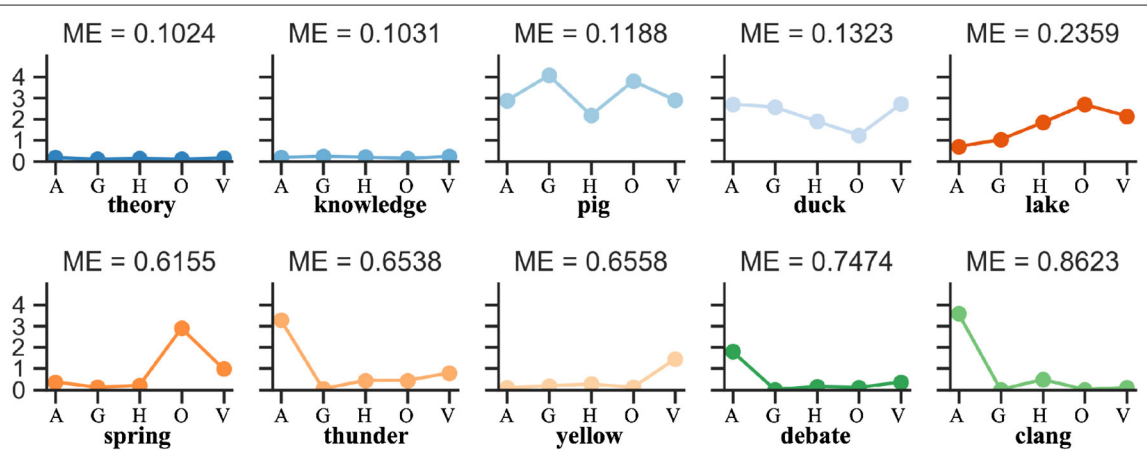


FIGURE 3 | Modality exclusivity demonstration. Modality exclusivity (ME) is a metric measuring how much of a concept is perceived through a single perceptual modality. For each concept, the value of ME is calculated as the perceptual strength range divided by the sum, and spanning from 0 to 100% for completely multimodal to completely unimodal perception.

We used the evaluation datasets WordSim353 and SCWS1994 for testing, and the inputs of the models were from different sources of multisensory representation datasets: BBSR, LC823an, Lancaster40k, and tested using two paradigms, IM and AM, respectively. For the AM paradigm, Izhikevich's seven models and LIF model were used, while for the IM paradigm, IF model were used for postsynaptic neurons and Izhikevich's seven models and LIF model were used for presynaptic neurons. The running time of all the tests is 100 ms and $I_{boost} = 100$.

4.1.2. Results and Analysis

From the overall results for both IM and AM paradigms, the integrated vectors are closer to humans than the original vectors based on our models: 37 submodels achieved better results for a total of 48 tests for both IM and AM, as **Table 2** shows. In terms of overall dataset, 15/16 tests work better for the BBSR dataset, 14/16 tests work better for LC823an, and 8/16 tests work better for Lancaster40k.

In almost all experiments, multisensory integrated representations based on our framework outperform unintegrated ones, with the exception of the instability shown in IM and AM paradigms when Lancaster40k is employed as the input. For any of the multisensory vectors, an integration way could be found to improve their representations.

4.2. Comparisons Between IM and AM Paradigms

Unlike the analysis of the macro-level above, in this section we introduce the concept feature norms to compare IM and AM paradigms from the micro-level perspective of each concept. Concept feature norms are a way of representing concepts by using standardized and systematic feature descriptions that mirror human comprehension. The similarities and differences of concepts are related to the intersection and difference of concept feature norms. McRae's concept feature norms, introduced by McRae et al. (2005), are the most prominent work in this area. They not only supplied 541 concepts with feature norms, but also proposed a methodology for generating them. For example, the feature norms of the concept "basement" are *"used for storage," "found below ground," "is cold," "found on the bottom floor," "is dark," "is damp," "made of cement," "part of a house," "has windows," "has a furnace," "has a foundation," "has stairways," "has walls," "is musty," "is scary," and "is the lowest floor."* Another semantic feature norms dataset analogous to McRae is CSLB (Centre for Speech, Language, and the Brain). They collected 866 concepts and improved the feature normalization and feature filtering procedure (Devereux et al., 2014). The McRae and CSLB criteria for human conceptual cognition are used in this research to investigate how each concept is similar to human cognition.

We compare and analyze IM and AM paradigms from two perspectives. First, we use the perceptual strength-related metric Modality Exclusivity to compare the two paradigms to explore the sensitive of them to the concepts' strength distribution of multisensory information. Then, to assess the generality of the IM and AM paradigms, we introduce nine psycholinguistic dimensions derived from the concept's nature, which are unrelated to perceptual strength.

TABLE 3 | The sensibility of IM and AM results to modality exclusivity.

Izhkevich model	AM		IM	
	McRae	CSLB	McRae	CSLB
RZ	0.0149	−0.0987	−0.1524	−0.4848
FS	0.2679	0.0901	−0.134	−0.4447
IB	−0.0559	0.0191	−0.2672	−0.4986
LTS	0.2113	0.035	−0.12	−0.0453
RS	0.1943	−0.0087	−0.006	−0.1997
CH	0.0988	0.0197	0.0294	0.0964
TC	0.2078	0.0398	−0.2115	−0.4761

4.2.1. Modality Exclusivity

Modality Exclusivity (ME) is a metric measuring how much of a concept is perceived through a single perceptual modality (Lynott and Connell, 2013). For each concept, the value of ME is calculated as the perceptual strength range divided by the sum, and spanning from 0 to 100% for completely multimodal to completely unimodal perception. **Figure 3** show some examples.

In the concept feature norms dataset, we first obtain all similar concepts $c^{similar}$ for each concept c based on the number of feature overlaps and record their rank list $R_c^{similar}$ sorted by similarity. Then, for all concepts, the corresponding multisensory integrated binary representations B^{IM} and B^{AM} are produced using the IM and AM paradigms, respectively. Next, for concept c, its k similar concepts $c_{IM}^{k\ similar}$ and $c_{AM}^{k\ similar}$ are computed based on integrated binarycodes and harming distance, respectively. We query the rank of these k similar concepts in the feature norms space $R_c^{similar}$ and take the average value, denoted as $kAR_{c_{IM}}$ and $kAR_{c_{AM}}$, which reflects the closeness of the multisensory representations to human cognition using two ways of integration in our framework. Smaller values of kAR indicate closer to human cognition at the microscopic level. Finally, we focus on all concepts in the representation dataset and calculate the correlation coefficients between the $kAR_{c_{IM}}$ or $kAR_{c_{AM}}$ arrays obtained using the above approach and the ME arrays corresponding to the concepts. This coefficient reflects the correlation between the two different multisensory concept integration paradigms and modal exclusivity. And in this experiment we only test the Izhikevich model and set k to 5.

The results in **Table 3** reveal the difference between IM and AM paradigms. The IM paradigm has a stronger negative correlation in both concept feature norms test sets, but the AM paradigms has a slightly positive correlation. We investigate this discrepancy further by viewing the FS model in detail, as shown in **Figure 4**. The results reveal that for concepts

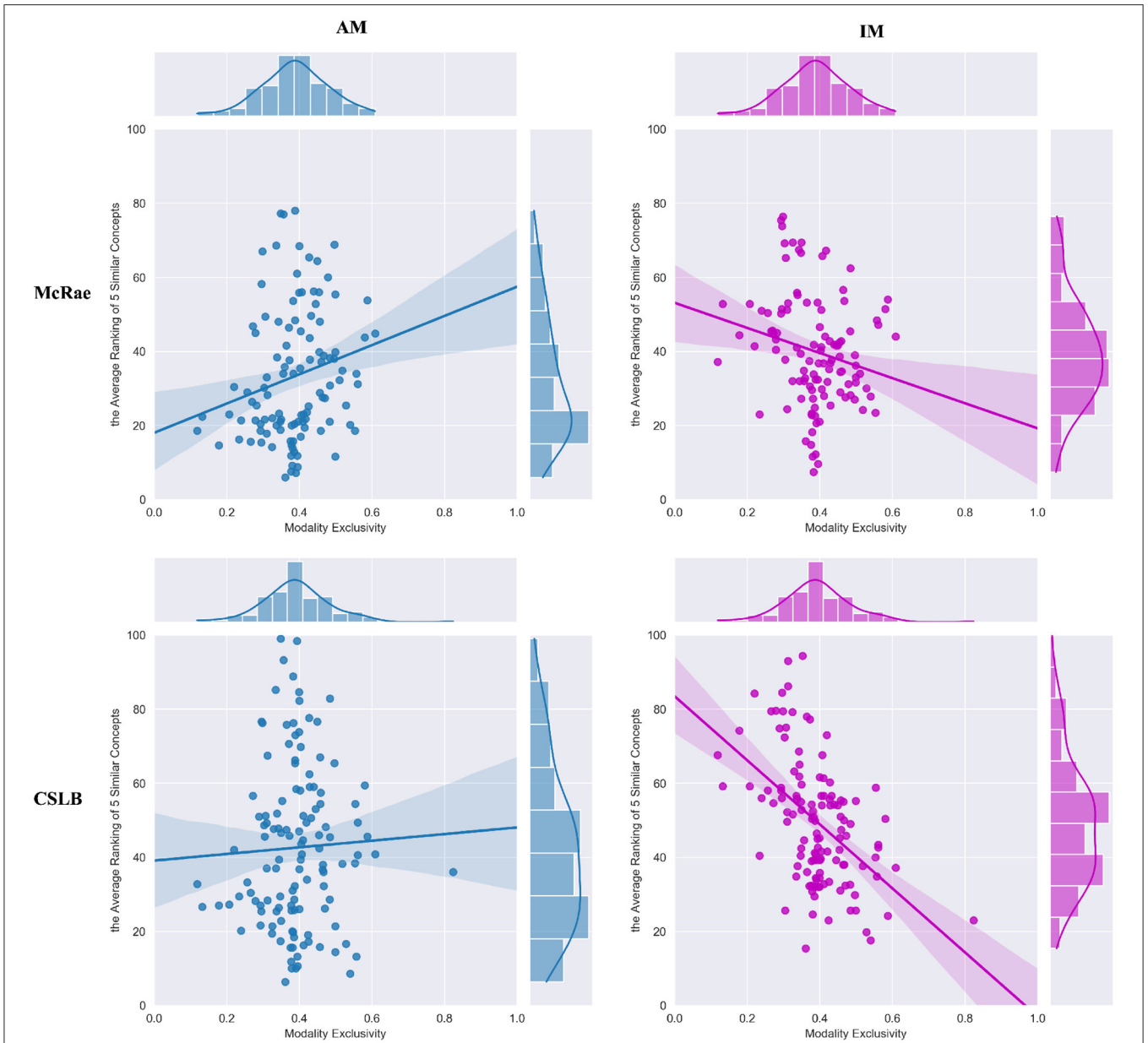


FIGURE 4 | The correlation between ME and average of five similar concept rankings.

with higher ME (such as "spring," "thunder," "yellow," "debate," "clang" in **Figure 3**), the IM paradigm is better at multisensory integration. While the AM paradigm is less input biased for each modality, it benefits the concept of uniform modal distribution (such as "theory," "knowledge," "pig," "duck," "lake" in **Figure 3**).

4.2.2. Generality Analysis

The ME metric used in the previous experiments is a perceptual strength-related indicator for the concept representation. In this part, we will test the framework from the input concept itself. And we introduce Glasgow norms which are a set of normative assessments on nine psycholinguistic dimensions: arousal (AROU), valence (VAL), dominance (DOM), concreteness (CNC), imageability (IMAG), familiarity (FAM), age of acquisition (AOA), semantic size (SIZE), and gender association (GEND) for 5,553 concepts (Scott et al., 2019).

In the same manner as the previous experiment. In concept feature norms, we first record all similar concepts for each concept, then sort them by similarity and rank them. Then, for IM and AM paradigms, we use the same concept input, get the integration vector for each concept, find their k similar, and get the mean value of their ranking in concept feature norms as $kAR_{c_{IM}}$ and $kAR_{c_{AM}}$. Finally, we determine the correlation

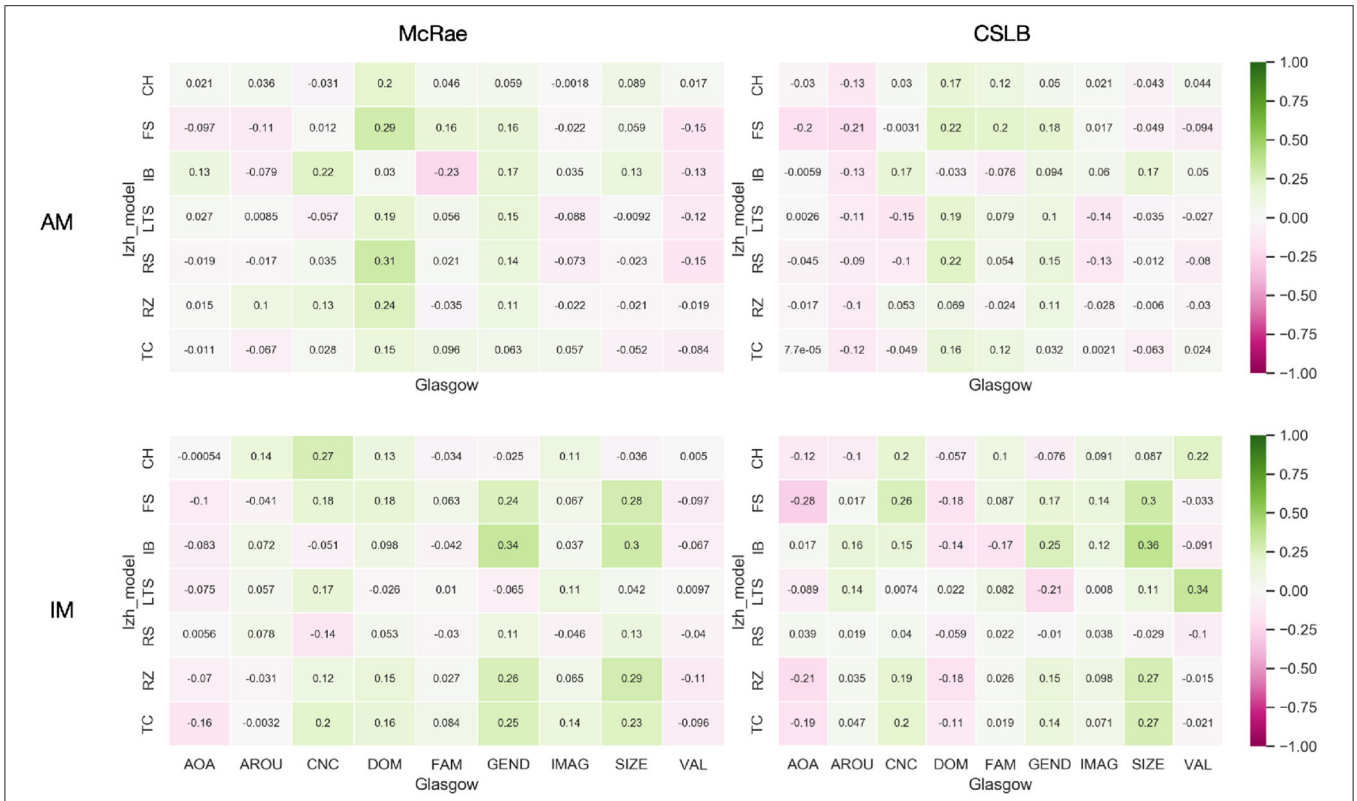


FIGURE 5 | The heatmap of generality analysis results.

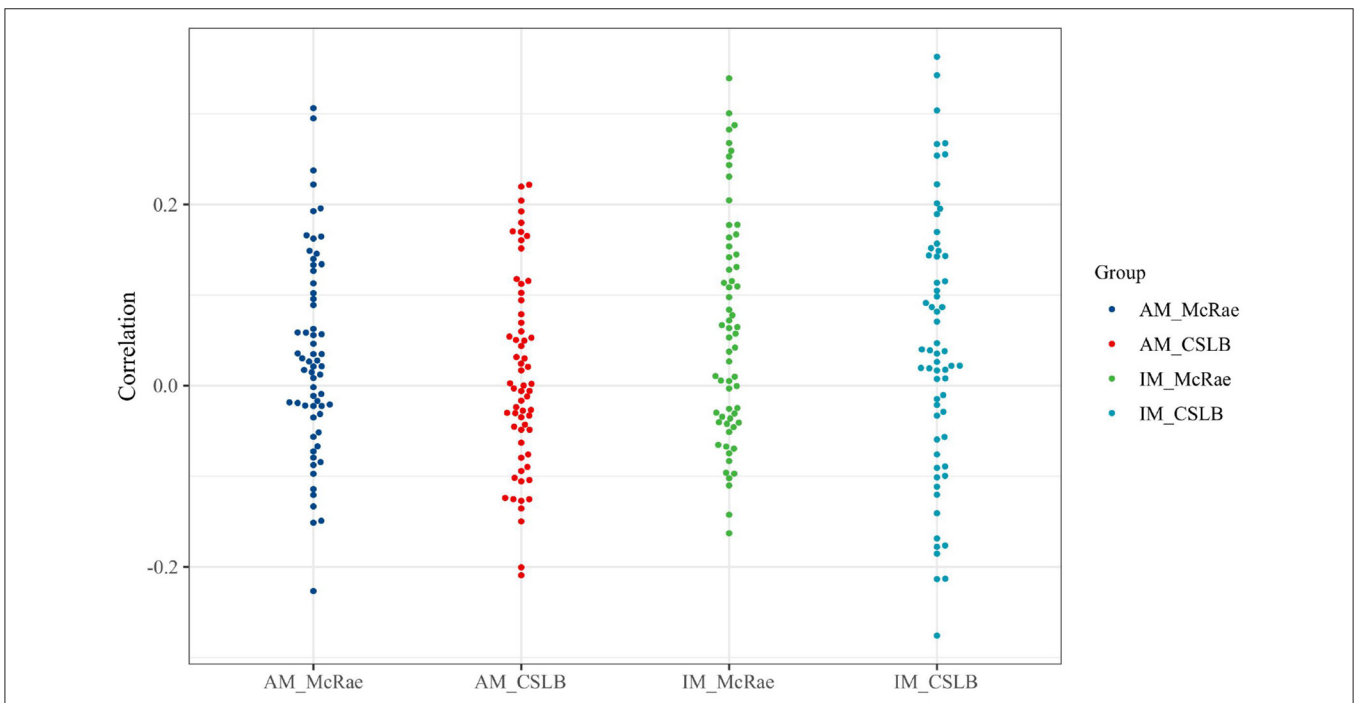


FIGURE 6 | The beeswarm of correlation distribution.

coefficient between each psychological characteristic and the concept's average ranking value *kAR* for the two paradigms. We still only test the Izhikevich model in this experiment, and the value is set to 5.

We used heatmaps (**Figure 5**) to visualize the correlation coefficients between the IM and AM paradigms' *kAR* and nine psycholinguistics in the two concept feature norms sets McRae and CSLB. Additionally, we omit the adopted Izhikevich submodels and provide the correlation coefficients using a beeswarm (**Figure 6**) to explain them more clearly.

According to the experimental results presented, the absolute values of all correlation coefficients are <0.3. The effect of vectors after integration of either IM or AM paradigms does not have any relationship with the nature of the concepts for several dimensions, including AOA, AROU, FAM, IMAG, and VAL. This indicates that both paradigms have good generality and the framework is not affected by the concepts themselves.

5. DISCUSSION

In this study, we propose a SNN-based concept learning framework for multisensory integration that can generate integration vectors based on psychologist-labeled multimodal representations. Vision, hearing, touch, smell, and taste are among the five modalities used in our research, which also includes a brain-like SNN model. We intend to add more brain-like processes in the future, such as multisensory fusion plasticity. The multisensory data we currently use are labeled by cognitive psychologists, which is relatively expensive and small, and in the future we consider expanding the relevant dataset by mapping for larger scale experiments. The current research focuses on multisensory representation of concepts, which is a subset of pattern representation in AI, and future research can be deeply integrated with downstream tasks to create AI systems that incorporate multisensory integration. At the same time, this places more demands on multisensory perceptrons. Human perception of concepts has not only multisensory perception but also more textual information based on abstract information, and it is also worth exploring how to combine these two parts to build human-like concept learning systems in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://osf.io/7emr6/; http://www.neuro.mcw.edu/resources.html; https://link.springer.com/article/10.3758/BRM.41.2.558; https://link.springer.com/article/10.3758/s13428-012-0267-0.

AUTHOR CONTRIBUTIONS

YW and YZ designed the study, performed the experiments, and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070100).

ACKNOWLEDGMENTS

We thank Dr. Yanchao Bi and Dr. Xiaosha Wang for helpful discussions and generous sharing of psychology-related researches.

REFERENCES

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pacsca, M., and Soroa, A. (2009). "A study on similarity and relatedness using distributional and word net-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Boulder, CO: Association for Computational Linguistics), 19–27. Available online at: https://aclanthology.org/N09-1003

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2014). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* 12, 1165–1187. doi: 10.1162/089976600300015547

Bi, G.-Q., and Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., et al. (2016). Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* 33, 130–174. doi: 10.1080/02643294.2016.1147426

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.* 49, 1–47. doi: 10.1613/jair.4135

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Cappe, C., Rouiller, E. M., and Barone, P. (2009). Multisensory anatomical pathways. *Hear. Res.* 258, 28–36. doi: 10.1016/j.heares.2009.04.017

Collell, G., Zhang, T., and Moens, M. -F. (2017). "Imagined visual representations as multimodal embeddings," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31 (San Francisco, CA: AAAI).

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *Behav. Res. Methods* 46, 1119–1127. doi: 10.3758/s13428-013-0420-4

Gao, Y., Hendricks, L. A., Kuchenbecker, K. J., and Darrell, T. (2016). Deep learning for tactile understanding from visual and haptic data. *arXiv:1511.06065*. doi: 10.1109/ICRA.2016.7487176

Gepner, R., Skanata, M. M., Bernat, N. M., Kaplow, M., and Gershow, M. (2015). Computations underlying drosophila photo-taxis, odor-taxis, and multi-sensory integration. *eLife* 4:e6229. doi: 10.7554/eLife.06229

Gerstner, W., and Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press. doi: 10.1017/CBO9780511815706

Hill, F., and Korhonen, A. (2014). "Learning abstract concept embeddings from multi-modal data: since you probably can't see what I mean," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Cambridge, MA: EMNLP), 255–265. doi: 10.3115/v1/D14-1032

Hill, F., Reichart, R., and Korhonen, A. (2014). Multi-modal models for concrete and abstract concept meaning. *Trans. Assoc. Comput. Linguist.* 2, 285–296. doi: 10.1162/tacl_a_00183

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for ComputationalLinguistics, Vol. 1* (Jeju Island: Association for Computational Linguistics), 873–882.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Netw.* 14, 1569–1572. doi: 10.1109/TNN.2003.820440

Kiela, D., and Bottou, L. (2014). "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Doha: EMNLP). doi: 10.3115/v1/D14-1005

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*. doi: 10.18653/v1/P18-1209

Lynott, D., and Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behav. Res. Methods* 41, 558–564. doi: 10.3758/BRM.41.2.558

Lynott, D., and Connell, L. (2013). Modality exclusivity norms for 400 nouns: the relationship between perceptual experience and surface word form. *Behav. Res. Methods* 45, 516–526. doi: 10.3758/s13428-012-0267-0

Lynott, D., Connell, L., Brysbaert, M., Brand, J., and Carney, J. (2019). The Lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods* 1–21. doi: 10.31234/osf.io/ktjwp

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 547–559. doi: 10.3758/BF03192726

Parise, C. V., and Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.* 7:11543. doi: 10.1038/ncomms11543

Rieke, F., Warland, D., Van Steveninck, R. d. R., and Bialek, W. (1999). *Spikes: Exploring the Neural Code*. MIT Press.

Roshan, C., Barker, R. A., Sahakian, B. J., and Robbins, T. W. (2001). Mechanisms of cognitive set flexibility in Parkinson's disease. *Brain A J. Neurol.* 124, 2503–2512. doi: 10.1093/brain/124.12.2503

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., and Sereno, S. C. (2019). The glasgow norms: ratings of 5,500 words on nine scales. *Behav. Res. Methods* 51, 1258–1270. doi: 10.3758/s13428-018-1099-3

Shams, L., and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends Cogn.* 12, 411–417. doi: 10.1016/j.tics.2008.07.006

Silberer, C., Ferrari, V., and Lapata, M. (2013). "Models of semantic representation with visual attributes," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Sofia: Association for Computational Linguistics), 572–582.

Silberer, C., and Lapata, M. (2014). "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Baltimore: Association for Computational Linguistics), 721–732. doi: 10.3115/v1/P14-1068

Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *Elife* 8:e47314. doi: 10.7554/eLife.47314

Troyer, T. W., and Miller, K. D. (1997). Physiological gain leads to high isi variability in a simple model of a cortical regular spiking cell. *Neural Comput.* 9, 971–983. doi: 10.1162/neco.1997.9.5.971

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Netw.* 60, 141–165. doi: 10.1016/j.neunet.2014.08.003

Ursino, M., Cuppini, C., Magosso, E., Serino, A., and Pellegrino, G. D. (2009). Multisensory integration in the superior colliculus: a neural network model. *J. Comput. Neurosci.* 26, 55–73. doi: 10.1007/s10827-008-0096-4

Verma, S., Wang, C., Zhu, L., and Liu, W. (2019). "Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis," in *International Joint Conference on Artificial Intelligence* (Macao). doi: 10.24963/ijcai.2019/503

Wang, S., Zhang, J., and Zong, C. (2018a). "Associative multichannel autoencoder for multimodal word representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 115–124. doi: 10.18653/v1/D18-1011

Wang, S., Zhang, J., and Zong, C. (2018b). "Learning multimodal word representation via dynamic fusion methods," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orelans, LA).

Wang, X., Men, W., Gao, J., Caramazza, A., and Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron* 107, 383–393.e5. doi: 10.1016/j.neuron.2020.04.010

Xu, Y., Yong, H., and Bi, Y. (2017). A tri-network model of human semantic processing. *Front. Psychol.* 8:1538. doi: 10.3389/fpsyg.2017.01538

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*. doi: 10.18653/v1/D17-1115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

The Initial Weights in IM

Similar to what cognitive psychologists (Ursino et al., 2014) have done before, we assume that for the concept s and its each modality $i \in [A, G, H, O, V]$ representations, $p(x_i|s) \sim N(x_i; s, \sigma_i)$, where $N(x; \mu, \sigma)$ stands for the normal distribution over x with mean μ and standard deviation σ. They are conditionally independent from each other and by Bayes' rule,

$$\begin{aligned} p(s|x_A, x_G, x_H, x_O, x_V) &\propto p(x_A, x_G, x_H, x_O, x_V|s) \\ &\propto \prod_i p(x_i|s) = \frac{1}{\prod_i(\sqrt{2\pi}\sigma_i)} e^{-\Sigma_i \frac{(x_i-s)^2}{2\sigma_i^2}} \\ &\propto -\Sigma_i \frac{(x_i-s)^2}{2\sigma_i^2} \end{aligned} \tag{11}$$

The maximum-a-posteriori estimation for s is $\widehat{s} = \Sigma_i \frac{\frac{1}{\sigma_i^2}}{\Sigma_i \frac{1}{\sigma_i^2}} x_i$, where $\frac{1}{\sigma_i^2}$ reflects the reliability of each modality for the same concept s. In our IM schema, we regard normalized reliability as the initial weights between pre-synaptic neurons (describing each modality) and the post-synaptic neuron(for integration), i.e.,

$$w_i^0 = \frac{\frac{1}{\sigma_i^2}}{\Sigma_i \frac{1}{\sigma_i^2}} \tag{12}$$

where we can get each σ_i *via* psychologist-labeled multisensory datasets.

frontiers | Frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 18 May 2022
doi: 10.3389/fnsys.2022.788486

Does Machine Understanding Require Consciousness?

*Robert Pepperell**

Fovolab, Cardiff Metropolitan University, Cardiff, United Kingdom

This article addresses the question of whether machine understanding requires consciousness. Some researchers in the field of machine understanding have argued that it is not necessary for computers to be conscious as long as they can match or exceed human performance in certain tasks. But despite the remarkable recent success of machine learning systems in areas such as natural language processing and image classification, important questions remain about their limited performance and about whether their cognitive abilities entail genuine understanding or are the product of spurious correlations. Here I draw a distinction between natural, artificial, and machine understanding. I analyse some concrete examples of natural understanding and show that although it shares properties with the artificial understanding implemented in current machine learning systems it also has some essential differences, the main one being that natural understanding in humans entails consciousness. Moreover, evidence from psychology and neurobiology suggests that it is this capacity for consciousness that, in part at least, explains for the superior performance of humans in some cognitive tasks and may also account for the authenticity of semantic processing that seems to be the hallmark of natural understanding. I propose a hypothesis that might help to explain why consciousness is important to understanding. In closing, I suggest that progress toward implementing human-like understanding in machines—machine understanding—may benefit from a naturalistic approach in which natural processes are modelled as closely as possible in mechanical substrates.

Keywords: machine learning, consciousness, naturalism, understanding, brain modelling

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Yoonsuck Choe,
Texas A&M University, United States
Ricardo Sanz,
Polytechnic University of Madrid, Spain

***Correspondence:**
Robert Pepperell
rpepperell@cardiffmet.ac.uk

Received: *02 October 2021*
Accepted: *12 April 2022*
Published: *18 May 2022*

Citation:
Pepperell R (2022) Does Machine Understanding Require Consciousness? Front. Syst. Neurosci. 16:788486. doi: 10.3389/fnsys.2022.788486

INTRODUCTION

The human capacity for understanding is a complex phenomenon that can involve many cognitive processes such as learning, insight, reward, memory, recognition, and perception. To implement this phenomenon mechanically—that is, to create machines that understand in the same way that humans do—presents an extremely daunting challenge.

Significant progress has been made toward this goal in the field of machine learning. We now have systems that perform very well, and sometimes better than humans, in language processing tasks (Devlin et al., 2019; He et al., 2021), image classification tasks (Zelinsky, 2013; Yang et al., 2019), and in playing complex games (Silver et al., 2016). Even though these systems are very effective in some situations, questions remain about how robust and generalisable they are (Shankar et al., 2020) and to what extent they are truly capable of human-like understanding or whether they are just computational manifestations of the Clever Hans spurious correlation effect

(Lapuschkin et al., 2019). In the early twentieth century, a horse of that name was touted as being able to solve arithmetic problems but was later found to be responding to involuntary cues in the body language of its trainer (Pfungst, 1911). This concern is related to the long-standing problem of authenticity raised by John Searle's Chinese Room argument about whether artificially intelligent machines have semantic understanding of the data they are processing or whether they are "blindly" following syntactic rules (Searle, 1984).

This article addresses the question of what constitutes understanding in humans and how it compares to the kind of understanding that is currently being implemented in digital computers. Partially following Les and Les (2017), I draw a distinction between "natural," "artificial," and "machine" understanding, as set out in **Table 1**. Natural understanding is the kind that humans are capable of; it is instantiated in the physical substrate of our nervous systems, in particular in our brains, and is regarded as "authentic." I take it that this is the kind of understanding that we ultimately aim to implement in machines. Artificial understanding is a kind of understanding that is currently implemented in highly trained digital computers and is exemplified by natural language processors like BERT (Devlin et al., 2019) and image classifiers like AlexNet (Krizhevsky et al., 2017). For the reasons just given, this kind of understanding does not perform as well, and nor is it regarded as authentic as, natural understanding.

I will analyse examples of natural and artificial understanding to describe some of their key properties and then compare these properties in light of the challenge of producing machine understanding, defined here as natural understanding implemented in a mechanical substrate[1]. The analysis suggests that natural understanding is distinguished from artificial understanding by its property of consciousness and that machine understanding systems may require this property if they are to overcome the limitations of current artificial understanding systems. This leads to the formulation of a hypothesis about why the capacity for consciousness is advantageous to natural understanding.

With some exceptions (e.g., Yufik, 2013; Hildt, 2019) recent theorists have argued that it is not a requirement that computer-based systems are capable of consciousness or genuine semantic appreciation in order to understand (e.g., Anderson, 2017; Les and Les, 2017; Thórisson and Kremelberg, 2017; Dietterich, 2019). The primary goal of these theorists is to design machines that perform well in problem solving, object detection, recognition, and language processing tasks (Zelinsky, 2013; Yang et al., 2019). Indeed, based on the levels of performance in these tasks achieved with recent machine learning systems, which are not claimed to be conscious, there is justification for arguing that consciousness is *not* a necessary requirement for artificial understanding, at least in some cases. But if our goal is to create machine understanding, as defined here, then the requirements may be different. Here I consider in more detail what constitutes natural understanding.

NATURAL UNDERSTANDING

Understanding cannot be easily or precisely defined. It has several subtly different senses in English (Oxford English Dictionary) and interpretations can vary from field to field. But is generally taken to mean the ability to "grasp" or "see" how different parts relate to or depend upon each other (Grimm, 2011). In this section I aim to provide a fuller description of some of the key properties of understanding by reference to two concrete examples. To take first a simple example from the domain of natural language understanding, for each of these sets of three words find the fourth word that they have in common:

1. PRINT BERRY BIRD
2. FENCE CARD MASTER
3. CONTROL PLACE RATE

These are examples of the Remote Associates Test commonly used to evaluate cognitive processes such as creative potential, problem solving, divergent thinking, and insight (Mednick, 1968; Bowden and Jung-Beeman, 2003). Consider your train of thought as you find the solution. When you begin the task the three given words seem to form an unrelated sequence. You may feel a mild sense of tension or anxiety as you struggle to find the answer. You probably take each given word in turn and wait for it to trigger other words, jumping between the given words until you alight upon a new word that links all three. Having found the common word, the three given words seem to subtly change their meaning by association with the common word. They acquire a new relationship with each other while retaining their distinct identities. Once you have understood the connection between each set of words you may feel a sudden mild sensation of pleasure or relief[2].

To take a more involved example from the domain of art interpretation, consider the painting reproduced in **Figure 1** that was painted by Pablo Picasso in 1910. It is a typical example of the analytic cubist style, developed by Picasso and Georges Braque in the years before world war I and depicts an arrangement of everyday household objects. If you are unfamiliar with the visual language of cubism it may be very hard—even impossible— to understand what it depicts and it usually takes some training and practice to unpick the objects it contains from the seemingly abstract forms.

Now consider the image presented in **Figure 2**. This shows the same painting, but this time some of the objects have been outlined and labelled. If you study this painting (which is known as "Still Life with Lemons") and then return to **Figure 1** you should now be able to recognise at least some of the items it

[1]A mechanical substrate is taken here to be a system composed of electrical and mechanical components that is designed to enable the processing of understanding, such as a computer or robotic system, that can receive data as input and produce a readable output.

[2]The answer in each case is 1. BLUE, 2. POST, and 3. BIRTH. In the paper from which these examples are taken 10% of the participants tested were able to find the correct answer to 1 in less than 2 s, while only 1% were in the case of 2 and none were in the case of 3 (Bowden and Jung-Beeman, 2003).

TABLE 1 | Definitions of the three kinds of understanding referred to in this article.

Definitions of kinds of understanding	
Natural understanding	The human-like capacity for understanding that is instantiated in our neurobiology, in particular in our brains
Artificial understanding	The capacity for understanding that is implemented in machine learning algorithms as instantiated in digital computers
Machine understanding	The human-like capacity for natural understanding implemented in a non-human mechanical substrate

contains without the guidelines. Given more time and effort you should eventually be able to piece together the entire composition. Arguably, you will then have gained a greater understanding of the meaning of the painting. Perhaps this understanding dawns through a gradual analysis of the relations between objects and their position in space. Or perhaps it appears as a momentary flash of insight—sometimes referred to as an "Aha!" moment—that is accompanied by the feeling of relief or satisfaction associated with a sudden gain of information (Muth and Carbon, 2013; Damiano et al., 2021). Either way, a significant shift has taken place in your perceptual and cognitive faculties such that objects and relationships between objects that were previously absent are now present, despite the fact that you are looking at the same image.

FIGURE 1 | A reproduction of a painting by Pablo Picasso from 1910. ©Succession Picasso/DACS, London 2022.

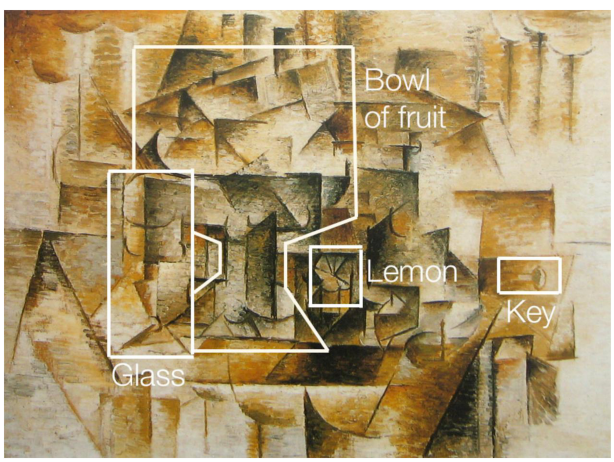


FIGURE 2 | A reproduction of *Still Life with Lemons* by Pablo Picasso from 1910 with outlined and labelled objects. The painting depicts a table containing a number of everyday household items, including glasses, a fruit bowl, a lemon, and a key. The edges and legs of the table can be seen to the left and right of the central grouping of objects.

What is going on at the perceptual, cognitive, and phenomenological levels during this acquisition of understanding? Prior to viewing **Figure 2** you probably experienced a more or less abstract array of patterns and marks, perhaps attended by a feeling of bewilderment or frustration. Then, using the outline guides provided in **Figure 2**, you began to separate the boundaries of certain objects from their surroundings until you established their individual identities and how they are spatially positioned in relation to each other and to the scene as a whole. According to the predictive coding theory of object recognition, your brain drew upon high-level cognitive models that influenced the processing of lower-level perceptual input via feedback in order to rapidly anticipate the most probable meaning of what is being perceived (Rao and Ballard, 1999). Once this meaning has been grasped you have created a new network of semantic associations around the image that are grounded in the wider context of your background knowledge and experience (Harnad, 1990).

Understanding, recognition, detection and learning are related but distinct processes. In one sense by studying this image you have learned to detect and classify or label the objects as any machine learning system might be trained to do with sufficient training examples and computer power. But in experiencing the phenomenal Aha! insight that accompanies the understanding you have not just produced a certain statistical output from a certain input; your perceptual, cognitive and phenomenological facilities have undergone a transformation from a state where that meaning is absent to one where it is present. There is evidence from brain imaging and behavioural studies that having undergone this experience with a small number of examples of cubist paintings people are able to recognise more objects more quickly in new examples while undergoing measurable differences in brain activation (Wiesmann et al., 2009)[3].

It is also important to stress that acquiring understanding does not merely entail local object detection and recognition but also in holding several distinct concepts in mind at once, along with each of their attendant associations, while forming a global conception of their interrelations and overall significance. These distinct concepts can be highly diverse, as is illustrated

[3]There is also evidence that learning to understand cubist paintings by recognising the objects in them increases people's aesthetic experience of the paintings (Muth et al., 2013).

in the cartoon by Saul Steinberg that featured on the cover of New Yorker magazine in 1969 showing the train of thought of a person viewing a cubist painting by Georges Braque (**Figure 3**)[4]. And they are not necessarily logically consistent. So, for example, a certain patch of painting composed of diagonal lines, curves and greyish-brown paint looks very unlike a lemon at the same time as being a lemon. This dichotomy between the material from which an image is constructed (paint, ink, pixels, etc.) and the objects that the material represents is a fundamental feature of all pictorial depiction (Pepperell, 2015), even if this cubist example is an extreme case of perceptual incongruence between the pictorial fabric and what is depicted. Yet despite this dichotomy we are rarely prevented from understanding that, when looking at a picture, a certain pattern of lines or colours simultaneously stands for a quite different object.

To summarise, these cases of problem solving and art interpretation demonstrate some of the key properties of natural understanding as broadly described here, namely that it is a form of reasoning, learning or recognition that is accompanied by a consciously experienced insight, motivated by a desire to overcome anxiety and gain pleasurable reward, that entails a diverse and sometimes contradictory set of associations, some of which depend on contextual knowledge and meaning prediction, that are bound together in a simultaneous cognitive state. These features are summarised in **Table 2**.

This list does not exhaustively describe each of the properties of natural understanding, nor does it collectively provide a precise definition. And it is worth noting that some forms of understanding are arrived at by a process of logical analysis rather than sudden insight (Jung-Beeman et al., 2004; Carpenter, 2020). But, at least with respect to the cases discussed here, this list is indicative of the range of properties that natural understanding entails. Assuming we can generalise from this to other cases of natural understanding, we have identified some of the properties that an authentic implementation of machine understanding would require.

[4]It is not clear from this illustration whether the collection of ideas and associations contained in the viewer's thought bubble are being experienced simultaneously or sequentially. Personal experience of studying artworks in this way suggests that it is probably a mixture of both.

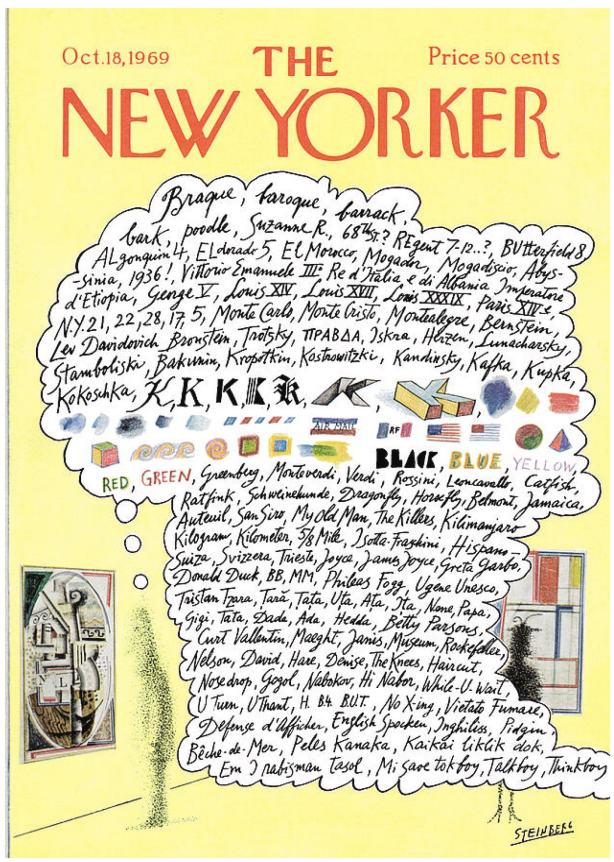


FIGURE 3 | Cover of New Yorker magazine with a cartoon by Saul Steinberg illustrating the diverse train of thought of a person viewing a cubist painting by Georges Braque.

ARTIFICIAL UNDERSTANDING

Having described some of the key properties of natural understanding we turn to the artificial kind as defined in the introduction. Many existing artificial intelligence systems are implemented in computational neural networks such as deeply layered convolutional neural networks that roughly approximate the function of neural cells in brain tissue. Contemporary deep neural networks evolved from early neurally inspired machine learning architectures such as the Pandemonium and the Perceptron pioneered in the 1950s (Rosenblatt, 1958; Selfridge, 1959). In these early models, continuous input data is first discretised by "feature detectors" and then passed to intervening layers of neurons that are weighted to respond to properties of the features. Based on the sum of all the weights the system reaches a decision processing about the most probable output. These models in turn inspired the later parallel distributed approaches to artificial intelligence that were developed by Rumelhart and McClelland (1986) and in many ways provided the core architecture of today's artificial neural networks and machine learning systems.

A typical artificial neural network tasked with, say, classifying objects in photographs will take an image as input, divide it into sub-sections (such as pixel colour values or clusters of pixels), pass those values to an array of nodes or neurons in one of what may be many interconnected "hidden" layers of such arrays, apply weights and biases in order to arrive at a probabilistic estimate of the likely class of the input, and pass the result to an output layer that can be read off by the user. By supplying the network with many training images, and by gradually optimising the weightings and bias using error correction techniques such as backpropagation, the network will eventually learn to classify its target objects with a degree of accuracy that depends on factors such as the size of the training dataset, the number of layers in the networks, and the amount of error correction provided. A simple feedforward example of this architecture is illustrated in **Figure 4**.

TABLE 2 | Summary of the key properties of natural understanding based on the cases of the remote associates task and the interpretation of a painting.

Key properties of natural understanding	
Insight	Aha! moment, or sudden change in how a stimulus is perceived entailing a revelation of new meaning that was previously absent
Reward	A positively valenced emotional state that intrinsically motivates effortful cognition
Learning	Adaptation by acquiring new knowledge that can be generalised to cases beyond the stimulus that produced the learning
Recognition	The ability to correctly classify a stimulus, or part of a stimulus, according to the features it presents or contains
Differentiation	The division of the perceptual stimulus into a multiple, diverse and sometimes contradictory set of meaningful elements
Integration	The unification of diverse perceptual elements into a single coherence experience, without diminishing their diversity
Context	Connecting to ideas, references and meanings that are not immediately present in the stimulus but are associated with it
Reasoning	A capacity to acquire new knowledge by logically inferring or extrapolating from existing data
Prediction	The ability to apply feedback from higher-level cognitive models to lower-level perceptual input to rapidly anticipate meaning
Consciousness	The state of being aware of the self and the environment, and in particular awareness of the stimulus and the response to it

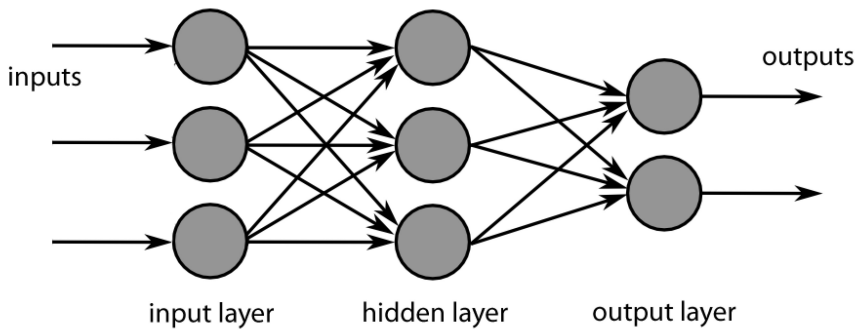


FIGURE 4 | A simple feedforward neural network architecture showing an input layer that serves to discretise the target data, one hidden layer that contains nodes or "neurons" that can adjust their probabilistic weights, and an output layer where the decision of the system can be read off.

Since the explosion of research in artificial neural networks and deep learning techniques in the 2010s, and the accompanying exponential increase in raw computing power, a plethora of designs and methods have evolved for implementing machine learning (LeCun et al., 2015; Aggarwal, 2018). In the case of a contemporary deep learning system like BERT, the Bidirectional Encoder Representations from Transformers, several methods are combined in order to optimise performance in a range of natural language understanding tasks, with the relative performance of different variants of BERT being tested against standardised benchmarks such as SuperGLUE (Wang et al., 2019).

In these tests, passages of text are presented to humans or computers to elicit a correct answer. Different kinds of understanding are tested, including reading comprehension, choosing correctly between alternatives, or reasoning correctly based on a hypothesis. For example, in the following causal reasoning task (Roemmele et al., 2011), given the statement: "My body cast a shadow over the grass" and the question: "What's the CAUSE for this?", the responder must choose between alternative 1: "The sun was rising" and alternative 2: "The grass was cut," the correct alternative being 1. In 2021, the DeBERTa variant of BERT was shown to surpass human performance against the SuperGLUE benchmark by a comfortable margin in some tests (He et al., 2021).

Image classification systems are designed to recognise, segment, or locate objects in images using convolutional neural networks that employ similar techniques to those of natural language processing systems but trained on vast databases of human annotated photographs stored on repositories such as ImageNet[5]. Competing models have been pitted against each other in contests such as the ImageNet Large Scale Visual Recognition Challenge or ILSVRC, which began in 2010 (Russakovsky et al., 2015). The ImageNet challenge uses a large dataset of annotated images from the database for training and a smaller subset for testing from which the annotations are withheld. The competing classifiers are required to perform several kinds of recognition and detection tasks on the test dataset, including predicting the classes of objects present in the image and drawing bounding boxes around objects (tasks not dissimilar to the cubist painting example discussed above). A breakthrough in image classification performance was made in 2012 with the introduction of the AlexNet architecture (Krizhevsky et al., 2017) which achieved the then unprecedented score in the ImageNet challenge of 63.3%. By 2021, systems such as Convolution and self-Attention Net (CoAtNet) were achieving accuracy scores of 90.8% (Dai et al., 2021).

Given that these natural language and image classification machines are routinely achieving 90 + % accuracy, and in some cases outperforming humans, there is a sense in which they can be rightly said to have a capacity for understanding, even though they are implemented in very different substrates

[5]http://www.image-net.org

from the biological tissue and processes that instantiates natural understanding. After all, show them a sentence with a missing word or a photograph containing many objects and they will reliably be able to predict the missing word or label the objects. This capacity for comprehension, reasoning, recognition, and detection implemented in digital computers is what is referred to here as artificial understanding.

The key properties of artificial understanding broadly described here are that it relies on training with large datasets through which the system learns by adjusting probabilistic weightings of the neurons, modified by error correction, resulting in statistical models that predict the most likely output for a given input, whether that is by detecting and labelling a class or reasoning from contextual data about the likely solution. To carry out this process input data is differentiated into parts and analysed to find patterns and associations between the parts which are then integrated to produce an output. These key properties of artificial understanding are summarised in **Table 3**.

Again, this is not a comprehensive list of the key features of nor a precise definition of artificial understanding. But on the basis of the natural language processing and image classification systems discussed here we are in a position to make some instructive comparisons between the natural and artificial kinds of understanding.

COMPARING NATURAL AND ARTIFICIAL UNDERSTANDING

As can be seen from **Table 4**, natural and artificial understanding, as described here, share several key properties, at least superficially, while some are unique to natural understanding. In this section, I compare these properties to establish how closely they are shared and what might be the significance of the differences.

Shared Properties

Prima facie, both kinds of understanding share some capacity for learning, recognition, differentiation, integration, utilisation of contextual information, reasoning, and prediction. These key properties are functionally similar in humans and artificial neural networks in that for certain tasks they can produce the same outputs from the same inputs, even if the substrates they are instantiated in and the ways they are implemented are very different. In the case of natural language processing, as noted, humans and computers can achieve comparable scores when assessed against the criteria used in the SuperGLUE tests, which are based on tests designed to measure reading ability, reasoning and comprehension skills in humans (e.g., Roemmele et al., 2011). Neural network-based image classification systems also now routinely equal and sometimes out-perform humans (Buetti-Dinh et al., 2019). And neuroscientific models of predictive coding in humans have inspired new designs of neural networks with enhanced object recognition capabilities (Wen et al., 2018). All this is testament to the remarkable proficiency of artificial understanding systems in emulating these human cognitive faculties.

Yet despite the impressive levels of performance achieved with some deep learning models, and their functional similarity with human capabilities, they still differ from and fall short of human-level performance in several ways, including in terms of how robust and generalisable they are. As noted above in the case of cubist painting interpretation, humans are adept at applying what they learn in one case to novel cases (Wiesmann et al., 2009). But because deep learning systems become very finely "tuned" to the limited datasets used to train them there is a danger of "shallow" learning, where the system's competences are limited to the training data and they are unable to adapt to new cases, as was shown recently in the domain of natural language inference (McCoy et al., 2019).

Meanwhile, image classification tasks using ImageNet-trained machine learning systems are yet to achieve human-level

TABLE 4 | Comparison between the key properties of natural and artificial understanding based on the cases discussed above.

Comparing properties of natural and artificial understanding	
Natural understanding	**Artificial understanding**
Learning	**Learning**
Recognition	**Recognition**
Differentiation	**Differentiation**
Integration	**Integration**
Context	**Context**
Reasoning	**Reasoning**
Prediction	**Prediction**
Consciousness	
Insight	
Reward	

Properties in bold are shared.

TABLE 3 | Summary of the key properties of artificial understanding based on the cases of natural language processing and image classification.

Key properties of artificial understanding	
Prediction	A capacity to estimate the correct output given a certain input based on probabilistic calculations
Learning	Improving performance of the system through a process of training and adaptation guided by feedback based on correctness of outputs
Differentiation	The division of the input into multiple features that can be analysed in terms of regularities and patterns
Integration	The summation of probabilistic analysis of the differentiated features to produce an output
Context	A table of statistical relationships that is extracted from the training data and used predict the most likely missing data
Recognition	Correctly identifying or labelling an object from a given input, or part of the input, by analysing its features and predicting the correct output
Reasoning	The capacity to select the correct conclusion given information that is implicit in the input but not explicitly stated

performance in certain tasks and are rated as being less robust and less generalisable than human agents (Shankar et al., 2020). The problems of robustness and generalisability in image classification algorithms were further highlighted by a study showing that the ability of leading models to understand the content of photographs was significantly impaired by difficult or "harder" cases, i.e., cases where the image content was more ambiguous (Recht and Roelofs, 2019).

The differences, or dissonances, between human and machine understanding (natural and artificial in the terminology used here) were explored by Zhang et al. (2019) in the context of Biederman's theory of human image understanding (Biederman, 1985). Biederman (1985) argued that image recognition depends upon first differentiating or segmenting the image into components that are invariant with respect to viewing position or image quality and from these components the understanding of the image as a whole is constructed. Zhang et al. (2019) asked both humans and neural network (NN) image classifiers to segment a set of images into "super pixels" that contained the portions of the image most salient to recognition. They found that humans and NNs tended to segment the image in different ways. When asked to recognise objects from the segmented portions only, NNs often out-performed humans on "easy" images, suggesting that humans and NNs were using different strategies to complete the task. But NNs performed less well than humans on more difficult or ambiguous images.

Collectively, this evidence suggests that while natural and artificial kinds of understanding do share the properties listed in bold in **Table 4**, at least at the functional level if not at the substrate level, and have comparable levels of performance in some cases, there are significant differences in how robust and generalisable they are and in how well they are able to deal with difficult cases. Moreover, questions remain about whether machine learning systems rely on spurious correlations—that they can be "right for the wrong reasons"—and whether they genuinely have a capacity for semantic appreciation. This leaves them vulnerable to Clever Hans and Chinese Room-style criticisms, viz., that they are not, by their essential nature, authentically cognising or understanding at all.

Unique Properties

The essential differences between natural and artificial understanding become more pronounced when we consider the key properties that are unique to natural consciousness, the most obvious being that it entails consciousness. Questions about the nature of consciousness, how it is instantiated in humans (or other creatures for that matter), and how it might be implemented in non-biological substrates are vast and deep and cannot be addressed in detail here. But it is necessary to briefly consider what the conscious property of natural understanding might be contributing to the phenomenon as a whole and why it might help to explain its essential difference from and advantages over the artificial kind. This is especially so given that two of the other key features of natural understanding as described here, namely insight and reward, are themselves aspects of conscious experience.

Consciousness can be defined as the state of awareness of self and environment, and while this begs the question of what is meant by awareness, I will take it that we are familiar with what it means in ourselves. One way to measure the difference between a system that is conscious and one that is not is that a conscious system such as a human brain displays very high levels of simultaneous differentiation and integration in its organisation and behaviour (Tononi et al., 1994). Of course, any system composed of different subsystems that are coupled together, i.e., a system of systems, will be differentiated and integrated to some degree (Nielsen et al., 2015). But in the case of the human brain this degree seems to be extremely large (Tononi et al., 1994) and far greater than in existing machine learning systems if we take the complexity of the system as a measure: it requires a convoluted neural network having seven layers to emulate the complexity a single human neuron (Beniaguev et al., 2021) and there are estimated to be around 86 billion such neurons and around the same number of non-neuronal cells in a human brain (Azevedo et al., 2009).

Recent evidence from the neuroscientific study of consciousness suggests that there is something particular about the way brain activity during conscious states is differentiated and integrated that contributes to the production of phenomenal states. The Global Neuronal Workspace Hypothesis (GNW) advocated by Baars et al. (2013) and Mashour et al. (2020) proposes a model of conscious processing in which localised, discrete and widely distributed cortical functions are integrated via reciprocally connected long-range axons. At any one time, information from one or more of these discrete functional processors can be selectively amplified and "broadcast" across the entire system, thus producing a single integrated, coherent experience for the conscious agent concerned. The Integrated Information Theory (IIT) of consciousness championed by Tononi and Koch (2015) and Tononi et al. (2016)—in some ways a competing theory to GNW—predicts that in order for a system such as a brain to be conscious it must display a high degree differentiation (by which they mean richness or diversity of information) and integration (by which they mean interdependence or interrelatedness of the information), the quantity of which is given by a value known as Φ. A fully conscious brain, for example, will contain a greater quantity of Φ than a partially conscious or unconscious brain.

Tononi and Koch point to work conducted by Casali et al. (2013) as empirical support for this hypothesis. By applying a magnetic pulse to the brains of people having varying levels of consciousness, including severely brain damaged patients showing little or no signs of conscious awareness, and then measuring the resulting patterns of activation using information-theoretical measures of complexity, the experimenters were able to reliably discriminate between levels of consciousness on the basis of how much differentiation and integration the patterns of activation displayed[6]. They found that greater levels of differentiation and integration reliably predicted higher

[6]The measure of complexity in this case was the compressibility (using the Lempel-Ziv algorithm) of the data generated by imaging the perturbation in the brains due to the magnetic pulse (Ziv and Lempel, 1977).

levels of consciousness, and could predict which people were unconscious when these levels fell below a certain threshold in their brains, such as in those with severe brain damage who were in a vegetative state. It is important to note that even though the brains of people with impaired consciousness were still functioning to some extent, and therefore displaying a high degree of differentiation and integration by the standards of many physical systems, they fell short of the threshold necessary to support full consciousness.

Further evidence that fully conscious states rely on maintaining a critical balance between activity in localised and segregated networks and globally integrated networks in the brain was provided by Rizkallah et al. (2019). Using graph-theory based analysis on high-density EEG data, the team showed that levels of consciousness decreased as the level of integration between long-range functional networks also decreased while, at the same time, information processing became increasingly clustered and localised. Besides disorders of consciousness, researchers have also shown that imbalances between local segregation and global integration in brain organisation are implicated in neuropsychiatric and other clinical disorders (Fair et al., 2007; Lord et al., 2017).

One difficult question raised by this evidence is whether there is a direct causal relationship between the levels of differentiation and integration observed in the activity of the brains of conscious people and their conscious states, or whether the correlation is spurious (Pepperell, 2018). The question is too philosophically involved to be addressed in depth here. But the phenomenal character of natural understanding, as described above, which entails an awareness of both the parts of the thing understood and the relations between the parts at the same time, is but one expression what seems to be a property of all conscious states, which is that they are experienced as simultaneously differentiated and integrated, as was observed by Leibniz (1998) in the eighteenth century and by many since[7]. Although this correlation is not proof of a causal link between phenomenology and underlying neurobiology, and nor does it explain why the particular kind or degree of differentiation and integration that occurs in conscious brains is critical, it does weaken any claim that the correlation is merely spurious.

With respect to the property of insight, which is consciously experienced, there is evidence from neuropsychology that comprehension or understanding, including that which is achieved through sudden insight or Aha!, is mediated by regions of the brain that are important for integration of differentiated brain processes (St George et al., 1999; Jung-Beeman et al., 2004). The same principle has been observed in the mechanisms that bind together widely distributed brain areas as object representations become conscious (Tallon-Baudry and Bertrand, 1999). Other studies have demonstrated that the appearance of sudden moments of insight or comprehension are in fact the culmination of multiple preceding brain states and processes, suggesting that insight favours the "prepared mind" and acts to draw these largely unconscious processes together into a single conscious state (Kounios and Beeman, 2009). This evidence therefore also points to a link between the underlying mechanisms that mediate consciousness and the phenomenology of natural understanding, or insight.

With respect to the property of reward, studies on the affective states of people who experience insights consistently show that they are emotionally diverse but positively valenced, with the most reported emotional states being happiness, certainty, calm, excitation, ease and delight (Shen et al., 2016). The affective states associated with insight and problem solving have been shown to depend on activity in regions of the brain associated with positive affect and reward and on task-related motivational areas as well as being implicated in processes of learning reinforcement, memory reorganisation, semantic coherence, and fast retrieval encoding (Tik et al., 2018).

The motivating power of potential reward, even when cued subliminally, was demonstrated by researchers who used a version of the remote associate task cited above to test problem solving performance in people (Cristofori et al., 2018). Based on their results they speculated that the potential for reward activated systems of the brain that reinforce behaviour, facilitate cognition, and enhance automatic integration of differentiated processes. The fact that they did so subliminally was argued to promote overall performance because cognitive resources were not diverted from conscious processes such as attention selectivity. Further evidence shows that mood can significantly affect a person's performance in problem solving, with people in positively valenced states of mind being able to solve problems or reach insights better than those in a less positive mood (Subramaniam et al., 2009). This finding reinforces the association between consciously experienced affect and capacity for understanding.

While is premature to draw firm conclusions from the neurobiological and psychological data relating to the key properties that are unique to natural understanding, it does seem to point toward a general trend: that the act of consciously understanding something is characterised by high degrees of simultaneous differentiation and integration—both neurobiologically and phenomenologically—and positively valenced affect that rewards problem solving and motivates learning. This comparative analysis between the shared and unique properties makes clear that although there are functional similarities between natural and artificial kinds of understanding there are also significant differences in function and in essence due, in part, to the conscious properties that natural understanding entails.

HYPOTHESIS

From the evidence and argument presented it is proposed that the present performance limitations of artificial understanding, and the questions about its authenticity noted in the introduction,

[7]Leibniz (1998) noted on several occasions that perception is "the expression of a multitude in a unity." More recently, Guilio Tononi, one of the prime movers behind IIT, stated: "consciousness corresponds to the capacity of a system to integrate information. This claim is motivated by two key phenomenological properties of consciousness: differentiation – the availability of a very large number of conscious experiences; and integration – the unity of each such experience" (Tononi, 2004).

may arise, at least in part, because it lacks the capacity for consciousness and the associated capacities for insight and reward that we find in natural understanding. This proposal can be expressed in the following hypothesis:

> *The capabilities deemed desirable but deficient in artificial understanding systems, viz., robustness, generalisability, competence in hard cases and authentic appreciation of meaning, occur in natural understanding, at least in part, because the motivation to gain insight, the unification of divergent concepts that the insight entails, and the reward that comes from achieving it are consciously experienced.*

The hypothesis suggests that there may be at least two reasons why the properties unique to natural understanding contribute to its capabilities and essential nature:

1. The promise of reward, and the positive affective states entailed by achieving reward, provide the system with the intrinsic motivation (Di Domenico and Ryan, 2017) to devote the necessary cognitive resources, such as memory search, object recognition, and selective attention, to the task in hand. This in turn reinforces learning and promotes memory reorganisation which improves performance in subsequent related tasks, particularly with respect to difficult cases, while also contributing to robustness.
2. The neurobiological activity that produces high degrees of simultaneous differentiation and integration, and which is associated with the occurrence of consciousness in humans, allows the understander to assimilate many diverse cognitive states into a single overarching cognitive state without effacing the differences between its constituent states. This neurobiological activity is reflected at the phenomenological level, as described in section "Natural Understanding," where natural understanding is characterised by the simultaneous "grasping" of diverse, and sometimes contradictory, concepts that form a meaningful conceptual whole.

Both of these reasons would require further analysis, investigation, and ideally empirical testing before we can draw any conclusions about their validity.

IMPLEMENTING MACHINE UNDERSTANDING

The question of how to implement machine understanding is related to, but distinct from, the question of how to implement machine consciousness (Haikonen, 2003; Pepperell, 2007; Yufik, 2013; Manzotti and Chella, 2018; Hildt, 2019). It is beyond the scope of this article to consider in any detail the conceptual and technical challenges that would face someone trying to encode the properties of natural understanding, as described here, in a non-human substrate. However, if we take it that it is the *natural* form of understanding that we are seeking to implement it follows that a naturalistic approach to creating such machines may be beneficial. By "naturalistic" I mean an approach that seeks to model the properties and functions of the naturally occurring phenomenon as closely as possible[8]. This would be in keeping with the early models of machine learning, cited above, that were directly inspired by natural biological processes.

Even though today's artificial neural networks are the direct descendants of these early naturalistically inspired models, they differ in important ways from the biological processes that underlie human cognition and consciousness. Consider, for example, that the adult human brain accounts for around 2% of body mass, but consumes around 20% of the body's energy budget when at rest, or some 20 W (Sokoloff, 1992; Laughlin, 2001). Yet while this might suggest that the brain is extremely energy hungry it is in fact extraordinarily efficient when compared to current day computers, especially those carrying out machine learning tasks (García-Martín et al., 2019). Training just one learning model just once can consume over 600,000 kWh (Strubell et al., 2019) while the amount of power (in terms of ATP availability) used by the cerebral cortex to carry actual computation has been estimated at around 0.1 W (Levy and Calvert, 2021).

Consider also that the organisation and exploitation of energy resources by the brain may be playing a far more significant role in the production of consciousness than is often assumed (Shulman, 2013). It can be argued that neuroscientific models of brain activity based primarily on digital information processing paradigms, which tend to predominate in the current literature, have underplayed the causal role of energy in the production of phenomenological states (Pepperell, 2018). For example, the groundbreaking work on measuring consciousness based on levels of differentiation and integration by Casali et al. (2013) noted above is commonly interpreted in information theoretical terms, where greater "information processing" relates to greater consciousness. Yet the same results could be equally well interpreted in energetic terms on the basis that greater levels of differentiation and integration of the metabolic processes in the brain are causally related to the greater levels of consciousness observed.

Recent attempts have been made to dramatically improve the energy efficiency of machine learning systems using neuromorphic hardware (Stöckl and Maass, 2021) and given the growing awareness of the environmental impact of machine learning computing this is likely to become a topic of more intense research (Dhar, 2020). Alongside this there is growing interest in better understanding the causal role that energy and work plays in mental functions like understanding (Yufik et al., 2017) and in thermodynamically inspired models of computing which attempt to harness the natural computational power of complex, self-organising, non-equilibrium systems (Hylton, 2020). At the same time arguments continue about whether the physical substrate in which any form of machine understanding or consciousness is implemented might have a critical bearing on its functionality and efficiency (Koene, 2012). Such arguments become especially relevant in the context of a naturalistic approach where, for example, the foundational role of energy acquisition and dissipation in artificial intelligence is highlighted

[8]For an example of a naturalistic approach applied to the problem of computationally modelling human visual space see Burleigh et al. (2018).

(Thagard, 2022). These developments suggest that considerations about the role that energy is playing in the natural system of the brain will increasingly inform future development of machine understanding and machine consciousness.

There is also an active line of research into designing systems capable of human-like faculties of perception, cognition and consciousness that is directly inspired by current neuroscientific theories of brain function (Marblestone et al., 2016). Prominent among these are models based on the Global Neuronal Workspace (GNW) theory cited above (Haqiqatkhah, 2019; Mallakin, 2019; Safron, 2020; VanRullen and Kanai, 2021). According to this theory, the brain contains many processes that are highly differentiated, localised, widely distributed and yet unconscious. Under certain conditions, these localised processes are broadcast across the entire brain network to form an integrated cognitive state which advocates of the theory argue is experienced consciously. Relating this theory to the example discussed in section "Natural Understanding," we could imagine the diverse perceptions, concepts, and associations generated by the cubist painting being instantiated in such distinct cortical processes across the brain. At the same time, the richly interconnected global workspace area containing long-distance axons is able to select one or more local processes to be broadcast to the entire system, thus allowing for widespread and simultaneous integration of the diverse processes, just as we experience when we have gained an understanding of the painting's meaning. Researchers such as VanRullen and Kanai (2021) have proposed methods for implementing the GNW in artificial neural networks with a view to improving the performance of current machine learning systems and potentially endowing them with a capacity for consciousness. If validated such brain-inspired machines would, in principle, satisfy the requirements for a mechanical implementation of natural understanding as defined here.

However, there are also reasons to be cautious about our ability to emulate natural understanding given the limitations of current computer architectures and therefore our ability to replicate natural processes in machines. A key property of the brain activity associated with consciousness is the presence of highly recursive neural processing in which activity is fed forward and backward throughout the brain, creating dynamic loops that bind local processes into larger global networks. GNW is one of several theories of brain function that foreground the importance of recursive, reentrant or recurrent processing (Edelman and Gally, 2013; Lamme, 2020) and diminution of such feedback activity has been shown to be one of the hallmarks of loss of consciousness during anaesthesia (Lee et al., 2009; Hudetz and Mashour, 2016). According to GNW, recurrent processing is one mechanism through which the simultaneity of conscious experience, in which multiple and diverse contents are bound into a single state of mind, is generated (Mashour et al., 2020). Given the highly complex physiological organisation of the brain, noted above, with its billions of interacting cells densely arranged in a three-dimensional lattice, it is not hard to appreciate how intricate multiscalar patterns of recurrent processing occur.

It is much harder to imagine how similar levels of recurrent processing could be implemented, or even simulated, in today's digital computer architectures. The physical design and operation of current computer hardware, which is generally controlled by a central processing unit that executes lines of computer code sequentially at a fixed clock rate, means that it is incapable of producing the highly non-linear and globally interconnected behaviour we observe among biological neurons. Moreover, the primarily linear nature of programme execution in current computers (notwithstanding parallel processing architectures) mitigates against the simultaneity of processing that seems to mark natural understanding and conscious processing. Of course, software-implemented feedback mechanisms are often integral to machine learning algorithms (Herzog et al., 2020) and neural feedback can be simulated in software (Caswell et al., 2016). Moreover, recent research into how recurrent processing in mammalian brains aids object recognition has also shown that it improves performance when simulated in neural nets (Kar et al., 2019). But generating the degree of recurrent and simultaneous processing necessary to support the synchronised integration of highly numerous and diverse modules, in the way that seems to mark understanding and consciousness in humans, may be far beyond the capability of current digital computer architectures given the requirement for complexity noted above.

This brief survey suggests that while natural biological processes continue to be a source of guidance and inspiration for those seeking to implement humans cognitive faculties such as consciousness in non-human substrates significant challenges and problems remain to be overcome.

CONCLUSION

This article addressed the question of whether consciousness is required for machine understanding. I have shown that although we lack a precise operational definition of understanding we can draw a useful distinction between the natural, artificial and machine kinds. By analysing concrete examples of natural understanding I have described some of its key properties and contrasted these with some of the key properties of artificial understanding. Although much more could be said about these properties and the contrasts between them, it is evident from the analysis presented here that the conscious properties of natural understanding mark a profound difference in both function and essence from artificial understanding, even though both share some functional similarities.

On the basis of this analysis, I have proposed a hypothesis that may help to explain the advantages that natural understanding has over the artificial kind, specifically in terms of its capacity for robustness and generalisability, its ability to deal with difficult cases, and in the authenticity of its cognitive and semantic processing. The practical challenges of implementing machine understanding have been briefly considered, and are clearly considerable. I suggest that a naturalistic approach to addressing this challenge may be beneficial, which means modelling the biological processes and structures that mediate understanding in humans and implementing these as efficiently as possible in a non-human mechanical substrate. However,

pursuing this approach may require us to move beyond today's computational architectures.

There are several limitations of the present study. To mention three: first, as stated at the outset, the phenomenon of natural understanding is highly complex and multifaceted, and we lack any precise definition of what understanding is. Worse, different people in different disciplines can take it to mean different things. As such, it is unlikely that any single analysis will be able to capture all its many psychological and neurobiological properties, define them all in detail, and explain how they all interact in a way that all agree upon. The pragmatic approach taken here has been to describe these properties in broad terms rather than define them precisely to provide a useful working account of the phenomenon so that it can be compared to other implementations of understanding in certain cases. But any future work in this area will inevitably require more precise and generally agreed definitions.

Second, the relationship between consciousness and understanding as discussed here is complicated by the fact that many of the cognitive processes that enable natural understanding occur subliminally, as noted above. Future investigations may need to take greater account of the role of unconscious processing in the brain, and how this might inform the design of machine understanding systems. This raises further questions about the extent to which we need to replicate natural brain processes and functions to successfully implement human-like capabilities in non-human substrates or whether designing machines that achieve more or less the same results, even if by very different means, will be sufficient "for all practical purposes" (Anderson, 2017).

Third, the problem of machine understanding is one that, to date and to a large extent, has been addressed within the discipline of computer science. The analysis presented in this article is highly interdisciplinary, drawing on knowledge from art history, psychology, neuroscience, computer science, consciousness studies and other fields. There is always a danger in such highly interdisciplinary studies of oversimplifying its constituent knowledge. However, the problem of machine understanding may be one that is so broad and so deep that we have no option but to take such a highly interdisciplinary approach. In which case we will need to establish protocols of cooperation among widely dispersed areas of research.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Aggarwal, C. C. (2018). *Neural Networks And Deep Learning: A Textbook*. Berlin: Cham Springer.

Anderson, M. (2017). *Why AI Works. Artificial Understanding*. Available online at: https://artificial-understanding.com/why-ai-works-b682a42b1ba3 (accessed Febuary 22, 2022).

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain *J. Comp. Neurol.* 513, 532–541. doi: 10.1002/cne.21974

Baars, B. J., Franklin, S., and Ramsoy, T. Z. (2013). Global workspace dynamics: cortical "binding and propagation" enables conscious contents. *Front. Psychol.* 4:200. doi: 10.3389/fpsyg.2013.00200

Beniaguev, D., Segev, I., and London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron* 109, 2727.e–2739.e. doi: 10.1016/j.neuron.2021.07.002

Biederman, I. (1985). Human image understanding: recent research and a theory. *Comput. Vis. Graph. Image Process.* 32, 29–73. doi: 10.1016/0734-189x(85)90002-7

Bowden, E. M., and Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behav. Res. Methods Instruments Comput.* 35, 634–639. doi: 10.3758/bf03195543

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., et al. (2019). Deep neural networks outperform human experts capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnol. Rep.* 22:e00321. doi: 10.1016/j.btre.2019.e00321

Burleigh, A., Pepperell, R., and Ruta, N. (2018). Natural perspective: mapping visual space with art and science. *Vision* 2:21. doi: 10.3390/vision2020021

Carpenter, W. (2020). "The aha! moment: the science behind creative insights," in *Toward Super-Creativity - Improving Creativity In Humans, Machines, And Human - Machine Collaborations*, ed. S. M. Brito (Intech Open: London). doi: 10.5772/intechopen.84973

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5:198ra105. doi: 10.1126/scitranslmed.3006294

Caswell, I., Shen, C., and Wang, L. (2016). Loopy neural nets: imitating feedback loops in the human brain. *Tech. Rep.*

Cristofori, I., Salvi, C., Beeman, M., and Grafman, J. (2018). The effects of expected reward on creative problem solving. *Cogn. Affect. Behav. Neurosci.* 18, 925–931. doi: 10.3758/s13415-018-0613-5

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). CoAtNet: marrying convolution and attention for all data sizes. *arXiv [Preprint]*.

Damiano, C., Van de Cruys, S., Boddez, Y., Król, M., Goetschalckx, L., and Wagemans, J. (2021). Visual affects: linking curiosity. Aha-Erlebnis, and memory through information gain. *J. Vis.* 21:2117. doi: 10.1016/j.cognition.2021.104698

Devlin, J., Ming-Wei, C., Lee, L., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Long and Short Papers), Stroudsburg, PA, USA.

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat. Mach. Intell.* 2, 423–425. doi: 10.1038/s42256-020-0219-9

Di Domenico, S. I., and Ryan, R. M. (2017). The emerging neuroscience of intrinsic motivation:a new frontier in self-determination research. *Front. Hum. Neurosci.* 11:145. doi: 10.3389/fnhum.2017.00145

Dietterich, T. (2019). *What Does It Mean For A Machine To "Understand"?*. Available online at: https://medium.com/@tdietterich/what-does-it-mean-for-a-machine-to-understand-555485f3ad40 (accessed February 21, 2022).

Edelman, G. M., and Gally, J. A. (2013). Reentry: a key mechanism for integration of brain function. *Front. Integr. Neurosci.* 7:63. doi: 10.3389/fnint.2013.00063

Fair, D. A., Dosenbach, N. U. F., Church, J. A., Cohen, A. L., Brahmbhatt, S., Miezin, F. M., et al. (2007). Development of distinct control networks through segregation and integration. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13507–13512. doi: 10.1073/pnas.0705843104

García-Martín, E., Rodrigues, C. F., Riley, G., and Grahn, H. (2019). Estimation of energy consumption in machine learning. *J. Parallel Distrib. Comput.* 134, 75–88. doi: 10.1016/j.jpdc.2019.07.007

Grimm, S. (2011). "Understanding," in *The Routledge Companion To Epistemology*, eds S. Bernecker and D. Pritchard (London: Routledge).

Haikonen, P. O. (2003). *The Cognitive Approach To Conscious Machines*. Exeter: Imprint Academic.

Haqiqatkhah, M. M. (2019). *Machine Consciousness and the Global Workspace Theory*. PhD Thesis, KU Leuven. doi: 10.31237/osf.io/vfy3e

Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.

He, P., Liu, X., Gao, J., and Chen, W. (2021). "DeBERTa: decoding-enhanced BERT with disentangled attention," in *Proceedings of 2021 International Conference on Learning Representations*, Ithaca, NY, Cornell University.

Herzog, S., Tetzlaff, C., and Wörgötter, F. (2020). Evolving artificial neural networks with feedback. *Neural Netw.* 123, 153–162. doi: 10.1016/j.neunet.2019.12.004

Hildt, E. (2019). Artificial intelligence: does consciousness matter? *Front. Psychol.* 10:1535. doi: 10.3389/fpsyg.2019.01535

Hudetz, A. G., and Mashour, G. A. (2016). Disconnecting consciousness: is there a common anesthetic end point? *Anesth. Anal.* 123, 1228–1240. doi: 10.1213/ANE.0000000000001353

Hylton, T. (2020). Thermodynamic computing: an intellectual and technological frontier *Proceedings* 47:23. doi: 10.3390/proceedings2020047023

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., et al. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biol.* 2:e97. doi: 10.1371/journal.pbio.0020097

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22, 974–983. doi: 10.1038/s41593-019-0392-5

Koene, R. A. (2012). How to copy a brain. *New Sci.* 216, 26–27. doi: 10.1016/s0262-4079(12)62755-9

Kounios, J., and Beeman, M. (2009). The Aha! Moment. *Curr. Dir. Psychol. Sci.* 18, 210–216.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Lamme, V. A. F. (2020). Visual functions generating conscious seeing. *Front. Psychol.* 11:83. doi: 10.3389/fpsyg.2020.00083

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* 10:1096. doi: 10.1038/s41467-019-08987-4

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Curr. Opin. Neurobiol.* 11, 475–480. doi: 10.1016/s0959-4388(00)00237-3

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lee, U., Kim, S., Noh, G.-J., Choi, B.-M., Hwang, E., and Mashour, G. A. (2009). The directionality and functional organization of frontoparietal connectivity during consciousness and anesthesia in humans. *Conscious. Cogn.* 18, 1069–1078. doi: 10.1016/j.concog.2009.04.004

Leibniz, W. G. (1998). *Discourse on Metaphysics, Section 9 (Loemker 1969: 308)*. Oxford: Philosophical Texts.

Les, Z., and Les, M. (2017). Machine Understanding - a new area of research aimed at building thinking/understanding machines. *Int. J. Math. Comput. Methods* 2:2017.

Levy, W. B., and Calvert, V. G. (2021). Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proc. Natl. Acad. Sci.* 118:e2008173118. doi: 10.1073/pnas.2008173118

Lord, L.-D., Stevner, A. B., Deco, G., and Kringelbach, M. L. (2017). Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. *Philos. Trans Royal. Soc. A* 375:283. doi: 10.1098/rsta.2016.0283

Mallakin, A. (2019). An integration of deep learning and neuroscience for machine consciousness. *Glob. J. Comput. Sci. Technol.* 19, 21–29. doi: 10.34257/gjcstdvol19is1pg21

Manzotti, R., and Chella, A. (2018). Good old-fashioned artificial consciousness and the intermediate level fallacy. *Front. Robot. AI.* 5:39. doi: 10.3389/frobt.2018.00039

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10:94. doi: 10.3389/fncom.2016.00094

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026

McCoy, T., Pavlick, E., and Linzen, T. (2019). "Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. doi: 10.18653/v1/p19-1334

Mednick, S. A. (1968). Remote associates test. *J. Creat. Behav.* 2, 213–214.

Muth, C., and Carbon, C.-C. (2013). The aesthetic aha: on the pleasure of having insights into Gestalt. *Acta Psychol.* 144, 25–30. doi: 10.1016/j.actpsy.2013.05.001

Muth, C., Pepperell, R., and Carbon, C.-C. (2013). Give me gestalt! Preference for cubist artworks revealing high detectability of objects. *Leonardo* 46, 488–489. doi: 10.1162/leon_a_00649

Nielsen, C. B., Larsen, P. G., Fitzgerald, J., Woodcock, J., and Peleska, J. (2015). Systems of systems engineering. *ACM Comput. Surv.* 48, 1–41. doi: 10.1002/9781119535041.part1

Pepperell, R. (2007). Applications for conscious systems. *AI Soc.* 22, 45–52. doi: 10.1007/s00146-006-0074-1

Pepperell, R. (2015). Artworks as dichotomous objects: implications for the scientific study of aesthetic experience. *Front. Hum. Neurosci.* 9:295. doi: 10.3389/fnhum.2015.00295

Pepperell, R. (2018). Consciousness as a physical process caused by the organization of energy in the brain. *Front. Psychol.* 9:2091. doi: 10.3389/fpsyg.2018.02091

Pfungst, O. (1911). *Clever Hans: The Horse Of Mr. Von Osten*. New York, NY: Henry Holt & Co.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Recht, B., and Roelofs, R. (2019). "Do imagenet classifiers generalize to imagenet?," in *Proceedings of the 36th International Conference on Machine Learning,* Long Beach, California, PMLR, 2019.

Rizkallah, J., Annen, J., Modolo, J., Gosseries, O., Benquet, P., Mortaheb, S., et al. (2019). Decreased integration of EEG source-space networks in disorders of consciousness. *Neuroimage. Clin.* 23:101841. doi: 10.1016/j.nicl.2019.101841

Roemmele, M., Adrian Bejan, C., and Gordon, A. (2011). "Choice of plausible alternatives: an evaluation of commonsense causal reasoning," in *2011 Proceedings of AAAI Spring Symposium Series*, Stanford, California, USA.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Foundations*. Cambridge, MA: MIT Press

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Safron, A. (2020). An integrated world modeling theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Front. Artif. Intell.* 3:30. doi: 10.3389/frai.2020.00030

Searle, J. (1984). *Minds, Brains and Science*. London: Penguin Books.

Selfridge, O. G. (1959). *Pandemonium: A Paradigm for Learning. In: Proceedings of the Symposium on Mechanisation of Thought Process: National Physics Laboratory*. London: Her Majesty's Stationary Office.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. (2020). "Evaluating machine accuracy on ImageNet," in *Proceedings of the 37th International Conference on Machine Learning,* Vienna, Austria, PMLR, 2020.

Shen, W., Yuan, Y., Liu, C., and Luo, J. (2016). In search of the "Aha!" experience: elucidating the emotionality of insight problem-solving. *Br. J. Psychol.* 107, 281–298. doi: 10.1111/bjop.12142

Shulman, R. G. (2013). *Brain Imaging: What It Can (and Cannot) Tell Us About Consciousness*. Oxford: Oxford University Press, doi: 10.1093/acprof:oso/9780199838721.001.0001

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Sokoloff, L. (1992). The brain as a chemical machine. *Prog. Brain Res.* 94, 19–33. doi: 10.1016/s0079-6123(08)61736-7

St George, M., Kutas, M., Martinez, A., and Sereno, M. I. (1999). Semantic integration in reading: engagement of the right hemisphere during discourse processing. *Brain* 122, 1317–1325. doi: 10.1093/brain/122.7.1317

Stöckl, C., and Maass, W. (2021). Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes. *Nat. Mach. Intell.* 3, 230–238. doi: 10.1038/s42256-021-00311-4

Strubell, E., Ganesh, A., and McCallum, A. (2019). "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. doi: 10.18653/v1/p19-1355

Subramaniam, K., Kounios, J., Parrish, T. B., and Jung-Beeman, M. (2009). A brain mechanism for facilitation of insight by positive affect. *J. Cogn. Neurosci.* 21, 415–432. doi: 10.1162/jocn.2009.21057

Tallon-Baudry, C., and Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn. Sci.* 3, 151–162. doi: 10.1016/s1364-6613(99)01299-1

Thagard, P. (2022). Energy requirements undermine substrate independence and mind-body functionalism. *Philos. Sci.* 89, 70–88. doi: 10.1017/psa.2021.15

Thórisson, K., and Kremelberg, D. (2017). "Do Machines understand? understanding understanding workshop," in *Proceedings of the 10th International Conference on Artificial General Intelligence (AGI-17), August 18*, Melbourne Australia.

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J., et al. (2018). Ultra-high-field fMRI insights on insight: neural correlates of the Aha!-moment. *Hum. Brain Mapp.* 39, 3241–3252. doi: 10.1002/hbm.24073

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. Royal Soc. B Biol. Sci.* 370:20140167. doi: 10.1098/rstb.2014.0167

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44

Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system *Proc. Natl. Acad. Sci. U.S.A.* 91, 5033–5037. doi: 10.1073/pnas.91.11.5033

VanRullen, R., and Kanai, R. (2021). Deep learning and the global workspace theory. *Trends Neurosci.* 44, 692–704. doi: 10.1016/j.tins.2021.04.005

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. (2019). "SuperGLUE: a stickier benchmark for general-purpose language understanding systems," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, eds H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, B. Fox, and R. Garnett (Vancouver, BC: Canada), 3261–3275.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., and Liu, Z. (2018). "Deep predictive coding network for object recognition," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm Sweden, 5266–5275.

Wiesmann, M., Pepperell, R., and Ishai, A. (2009). Training Facilitates Object Perception in Cubist Paintings. *Neuroimage* 47:S85. doi: 10.1016/s1053-8119(09)70634-2

Yang, M. Y., Rosenhahn, B., and Murino, V. (2019). *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*. Cambridge, MA: Academic Press.

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solit. Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M., Sengupta, B., and Friston, K. (2017). *Self-Organization in the Nervous System*. Lausanne: Frontiers Media SA.

Zelinsky, G. J. (2013). Understanding scene understanding. *Front. Psychol.* 4:954. doi: 10.3389/fpsyg.2013.00954

Zhang, Z., Singh, J., Gadiraju, U., and Anand, A. (2019). "Dissonance between human and machine understanding," in *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW)*, Ithaca, NY, Cornell University, 1–23. doi: 10.1097/HNP.0000000000000010

Ziv, J., and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory* 23, 337–343. doi: 10.1109/TIT.1977.1055714/

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Systems Neuroscience

REVIEW
published: 28 July 2022
doi: 10.3389/fnsys.2022.796100

An Expanded Framework for Situation Control

James Llinas[1]* *and Raj Malhotra*[2]

[1] *Industrial and Systems Engineering Department, University at Buffalo, Buffalo, NY, United States,* [2] *U.S. Air Force Research Laboratory Sensors Directorate, Wright-Patterson Air Force Base, Dayton, OH, United States*

There is an extensive body of literature on the topic of estimating situational states, in applications ranging from cyber-defense to military operations to traffic situations and autonomous cars. In the military/defense/intelligence literature, situation assessment seems to be the *sine qua non* for any research on surveillance and reconnaissance, command and control, and intelligence analysis. Virtually all of this work focuses on assessing the situation-at-the-moment; many if not most of the estimation techniques are based on Data and Information Fusion (DIF) approaches, with some recent schemes employing Artificial Intelligence (AI) and Machine Learning (ML) methods. But estimating and recognizing situational conditions is most often couched in a decision-making, action-taking context, implying that actions may be needed so that certain goal situations will be reached as a result of such actions, or at least that progress toward such goal states will be made. This context thus frames the estimation of situational states in the larger context of a control-loop, with a need to understand the temporal evolution of situational states, not just a snapshot at a given time. Estimating situational dynamics requires the important functions of situation recognition, situation prediction, and situation understanding that are also central to such an integrated estimation + action-taking architecture. The varied processes for all of these combined capabilities lie in a closed-loop "situation control" framework, where the core operations of a stochastic control process involve situation recognition—learning—prediction—situation "error" assessment—and action taking to move the situation to a goal state. We propose several additional functionalities for this closed-loop control process in relation to some prior work on this topic, to include remarks on the integration of control-theoretic principles. Expanded remarks are also made on the state of the art of the schemas and computational technologies for situation recognition, prediction and understanding, as well as the roles for human intelligence in this larger framework.

Keywords: stochastic control and time-varying systems, situation control, situation assessment, estimation, prediction

OPEN ACCESS

Edited by:
Rosalyn J. Moran,
King's College London,
United Kingdom

Reviewed by:
Patricia Dockhorn Costa,
Federal University of Espirito Santo,
Brazil
Robinson E. Pino,
Office of Science (DOE), United States

***Correspondence:**
James Llinas
llinas@buffalo.edu

Received: *15 October 2021*
Accepted: *12 April 2022*
Published: *28 July 2022*

Citation:
Llinas J and Malhotra R (2022) An Expanded Framework for Situation Control.
Front. Syst. Neurosci. 16:796100.
doi: 10.3389/fnsys.2022.796100

INTRODUCTION AND REVIEW OF CURRENT RESEARCH

The concept of a "situation" can be thought of as describing a portion of a real-world that is of interest to a participant in that portion of the world. An understanding of a situation is needed and useful toward guiding or assessing the need for possible action of the participant in that situation. Action of a participant may also be needed to possibly alter the situation if it is in an undesirable

state (assuming resources capable of affecting the situation are available), or for the participant to alter his position in the situation. For a human participant, the mental faculties of human cognition, such as consciousness (awareness), reasoning, formation of beliefs, memory, adaptation, and learning, frame the functional aspects of a process of cognitive situational understanding, related to the notion of sensemaking (see, e.g., Pirolli and Card, 2005; Klein et al., 2007).[1] Acting on the situation, however, leads to the process of cognitive situation control, as well described in various of Jakobson's papers (Jakobson et al., 2006, 2007, Jakobson, 2008; Jakobson, 2017) that, in part, motivated this work. A depiction of that process is shown in **Figure 1** taken from Jakobson (2017); we offer here an abbreviated description of that process. The cycle starts with the existence of some (real, true) condition in the world, shown here by Jakobson as the "Operational Theater" which, as shown, can be affected by nature (that is, a context affected/defined by various contextual factors) and possibly of hostile or adversarial agents. That real situation is observed by imperfect and often multiple, multimodal sensors, and possibly human observers to support an estimation process that yields a "recognition" of the situation (a state estimate) that may be reasoned over by a human agent, or that provides an input to a subsequent automated process. (The situational picture so derived is understood to be only a part of some larger situational construct.) Jakobson calls this estimate the "Abstract Situation" in **Figure 1**. Given the current, recognized situation derived largely from observation, a Situation Learning process evolves from what we will call a contextual learning or a model-building process that could also be called Situation Understanding. Such a process implies an ability to develop a generalized, broader conception from the particulars of the current recognized picture and exploiting contextual factors either known *a priori* or collected in real-time. This process is similar to Bruner's view that "mental modeling is a form of information production inside the neuronal system extending the reach of human cognition 'beyond the information given'" (Bruner, 1973). Following Jakobson, the recognized, learned situation is compared to a goal situation that presumably can be specified *a priori* or in real-time, and a difference is computed between the two situational states by a Situation Comparator function. That difference can be considered, from a control process point of view, as an "error" signal; if that (likely stochastic) difference is high enough (in consideration of an estimated state variance), actions need to be contemplated and assessed in a decision-making process, and once defined are enabled onto the current situation in an effort to "move" the situation toward the goal state. Note that Actions or Effects on to the situation can only be realized through whatever "Actuators" or Resources may be available to this control process.

There are two classes of "Resources" in this characterization: Observational Resources and "Actuators" or Resources that can enable changes in the real situation; these could also be called "Effectors." The effective design of a process of managing these resources raises some challenges. For the Observational Resources, they first of all have to support the process that forms a recognized situational picture, possibly in the face of the "Five V's" of modern-day Big Data environments,[2] since this process does not start without an (estimated) recognized situational picture. To the extent then that the Observational Resources are a fixed resource set, and have any slack in their employment, they can also be used/multiplexed to support the employment of Effectors, as Effectors will need to be directed in some way. We submit that there is a time delay of possibly widely varying extent between the time of (initial) Situation Recognition and the eventual time of action of the Effectors; that is, most situations are continually unfolding and changing; they are dynamic. This being the case, it can be that there is a meaningful difference between the initial recognized situation and the situation that is eventually acted upon; such differences may result in very incorrect results of Actions if not accounted for. Thus, we assert that there will usually be a need for a Situation Prediction capability to create a temporal synchronization in this control process by propagating the situational estimate to the (expected, estimated) time of action. Then, just before acting, the predicted situation should be verified, this also requiring Observational resources. In sum, the Observational Resources will be shared over three different functional operations, as follows:

- Synchronizing Observation to Situational Velocity, Volume, Variety, Veracity, and Value in support of Situation Recognition
- Observation Multiplexing to support employment of Resources/Effectors
- Observation Multiplexing to support Situation Prediction confirmation.

A factor that will be very important in determining the process context for Situation Management and Control is the assessed rate at which the situation is unfolding; that is, the Operational Tempo ("OpTempo") of the situation. This factor needs to be weighed in relation to both the scanning/sampling rate of the Observational Resources, the prediction interval, sensor resolution factors, and in fact the viability of the overall process; if the situation is unfolding at a rate faster than it can be feasibly observed, forming dependable situation estimates will be very difficult, and situational predictions will be equally hard. This balance changes the dependence of the Learning/Understanding process between *a priori* knowledge and real-time observational data; uncertainties in the consequent estimated situation will also be affected. Estimating situational OpTempo should therefore be a fundamental requirement of the Situation Recognition function, as it is a critical process design and management parameter, setting the overall "clock" for this control process. The notion of OpTempo is in the fashion of a meta-metric, since any situation will be comprised of multiple component processes unfolding at varying rates. Note too that there are optimization issues lurking here, as regards defining how optimal

[1] Sensemaking is not the same as understanding; sensemaking involves interplay between foraging for information and abstracting the information into a representation called a schema that will facilitate a decision or solution (http://www.peterpirolli.com/Professional/Blog__Making_Sense/Entries/2010/8/16_What_is_sensemaking.html).

[2] Volume, velocity, variety, veracity, and value.

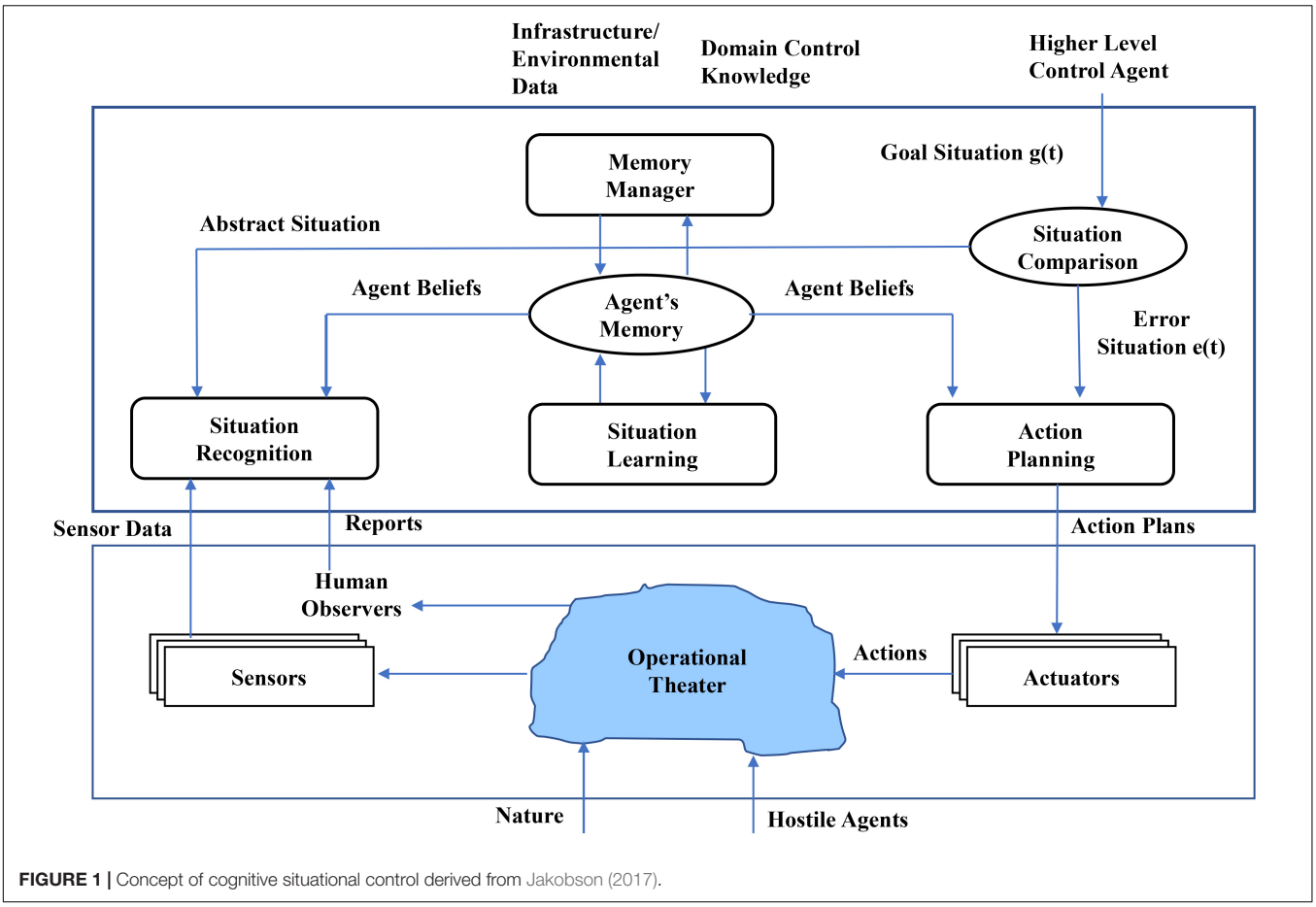


FIGURE 1 | Concept of cognitive situational control derived from Jakobson (2017).

co-employment of bounded Observational Resources will be managed across these process needs.

Jakobson does not elaborate on the functions of Situation Learning nor on the Memory-based processes shown in **Figure 1** (by choice, deferring those topics to future publications). He does elaborate on the functions of Situation Recognition as a tree-like hierarchical structure of component situation recognition sub-processes. A disaster-based use case is described within which an action-taking process that is also layered is elaborated. Jakobson, along with others on various occasions, has produced a number of papers on the central themes of cognitive situation management and many related topics that bear on the overarching topic of situation management (see prior citations and Jakobson et al., 2006; Jakobson, 2008).

To provide a historical perspective related to the process of situation control, we cite here the work of John Boyd, a military strategist and United States Air Force Colonel who in the 1980's put forward the paradigm that has come to be called the "OODA Loop," OODA an acronym for Observe-Orient-Decide-Act (see Boyd, 1986), but there are many papers, and a wide range of publications related to this paradigm if one searches on the web. It should be clear that these functions are quite similar to those depicted in **Figure 1**, with "Orient" perhaps needing clarification. Before remarking on Orient, it is emphasized that the OODA process was framed as a mental process, and then was studied by many to expand the framework to a potentially computational basis. Orient then was about mental modeling that built a mental model of a situation by consideration of prior knowledge (long-term memory), new information, cultural factors (a contextual effect), and other factors. This situation control type paradigm has found its way into business intelligence settings, game theory, law enforcement, and a multiplicity of other applications. A thorough review of the OODA process is provided in Richards (2012), although there are many publications about this process that addresses situation control.

Our intent in this paper is to expand the framework of cognitive control in terms of our views of several other component processes (forthcoming), and in discussing these additional processes, to relate them to research and capabilities in the cognitive neurosciences and machine understanding domains.

SITUATION CONTROL IN CRISIS MANAGEMENT

There is a large literature on crisis management and disaster management. In many cases, the characterization of the process begins with an assumption that certain of these problems can be anticipated, since in many cases an assessment of vulnerability to

specific types of crises can be analyzed, such as in the cases of natural disasters. The ability to achieve Situation Recognition in these cases benefits from recognizing *anticipated* early signals of the onset of the event, among other factors. However, there are many other crises that do not follow this model, either because they are of a rare type or perhaps because they are perpetrated by some actors; situation recognition in these cases is both more difficult and will also take more time for evidence accrual. Perpetrated crises are analogous to military-type crises and can have similar properties such as the employment of deception techniques and other complications; these factors re-orient the situation assessment process to one of adversarial reasoning. In any setting involving situation state estimation, an early question has to do with whether the setting is a natural one where phenomena are driven by natural causes or whether the setting comprises a two-sided, adversarial context. The case involving adversaries can be related to the case of "Information Warfare," (IW), where the two sides are manipulating information, the bases for perception and inference, to their advantage. The larger purpose of these operations is to manage adversarial perceptions by structuring the information available to an adversary to be compliant with that perceptual construct. Another topic related to deception is denial of information by covertness, camouflage, jamming, and other means. Deception and denial strategies work because of exploitation of reasoning errors, cognitive limitations, and cognitive biases (Elsaesser and Stech, 2007). The most important errors are:

- Reasoning causally
- Failure to include a deception hypothesis
- Biased estimates of probabilities
- Failure to consider false positive rates of evidence.

In our own experience in dealing with an earthquake disaster case, there was the additional complication of multi-jurisdictional participants, all taking different views of the integrated situation and what resources are to be deployed and controlled. This latter case involved additional processes of consensus-forming and complex communications to both recognize and predict situational states. In our disaster example and in most crisis problem contexts, a top priority is life-saving and casualty recovery, and the situation to both understand and control is that which relates to all of the dimensions of casualty-recovery operations. Such operations are dependent on vulnerable infrastructure components such as airports, ambulance depots, and electrical power. In addition, it is very typical in crises that there are cascading effects; in the case say of an earthquake, the tremors will cause primary problems such as building collapses but will in addition rupture gas lines leading to fires as secondary threats. These same cascading events occur in other crises as well, such as in wildfires, where entry and exit routes are compromised by evolving fire patterns, and where wildfire observation such as from drones is affected by dense smoke patterns; all of these factors drive a need to model the dynamics of situation control patterns. A core challenge in all situation control problems is achieving synchronization of the situation recognition, prediction, and understanding processes with control-related and action-taking processes. That is, there are the issues of gaining situation awareness and maintaining situation awareness, while comparing situational conditions to those desired and subsequently deciding on specific control actions.

Related Work; A Sampling

As noted above, there is a lot of literature on crisis and disaster management for which the topic of situation control would seem to be of interest. Relatively few papers in this field, however, address end-to-end process issues and models in the systemic context of this paper, although there many papers that address portions of the entire process; we sample a few here.

For example, it is clear that any Situation Control process must also be managing data and might require ancillary analytical support operations. The paper by Hristidis et al. (2010) provides one overview of data management and analysis processes in a stressing disaster-type situation. Information extraction, retrieval, and filtering processes (similar to data preparation processes in data fusion operations) are needed to extract relevant data of satisfactory quality for subsequent operations. Aspects of the supporting process infrastructure are addressed here as well, such as the need for a consistently formatted data base. This work is focused on textual data (often called "soft" data to distinguish it from quantitative "hard" data from electromechanical sensors), an important class of data for situation control, often not addressed.

Zambrano et al. (2017) provide an interesting aspect of a modern-day situation control problem regarding the use of cellphones; most modern contexts where a situation is evolving will involve cellphones carried by many people, and cellphone data of various type can contribute to both the estimation of the situation and aid in controlling the situation. This paper interestingly brings together a detailed data fusion process model, following the well-known JDL Data Fusion Process Model (see Llinas et al., 2004) and builds an end-to-end situation estimation process model based on cellphone-captured data. The main contribution here is the messaging protocol for information exchange in complex cellphone networks, and support to early warning notifications in real time.

The paper by Van de Walle et al. (2016) provides some interesting views for enriching raw incoming information by adding a summary of the information received, and by channeling all incoming information to a central coordinator who then decides upon further distribution within the team. This paper is largely about information quality, a factor that is important in all information operations. In a manner similar to assigning "pedigree" to information sources based on analytical or experiential bases, this paper discusses notions of information richness that can be based on reputation or on analytical methods that compute metrics for information sources based on notions of completeness or timeliness, and other such quality-influencing factors. While information can be enriched in various ways, in this research "enriched information" is defined as information that combines information from different sources and is represented in a format with which professional crisis responders are familiar, similar to the association and

combining operations in a data fusion process. Information that is not aggregated nor represented in a specific format is considered "raw" or non-enriched. This work carries out a series of experiments to explore the hypotheses related to information enrichment and centralized decision-making, concluding that that enriched and non-enriched information conditions are significantly different only if information is centralized.

In Costa et al. (2012), a Situation Modeling Language (SML) is developed, which is a graphical language for situation modeling, and an approach to situation detection and recognition based on the SML model is realized by linking the model to a rule-based scheme. The motivation for this paper comes in part from a view of Kokar et al. (2009) that argues, from an ontological point of view, that "to make use of situation awareness [. . .] one must be able to recognize situations, [. . .] associate various properties with particular situations, and communicate descriptions of situations to others." In addition to supporting an ontological foundation related to anything having to do with situations, this paper has many features that resonate with our own ideas, for example in defining situations as composite entities whose constituents are other entities, their properties, and the relations in which they are involved. This leads to an approach which is similar to an ontological approach that we also argue for in this paper, and also to a graphical construct that we also support as the correct modeling basis for these problems. This work also concerns itself with formal semantics which are quite necessary for these problems since clear semantics aid in clarifying combinatoric complexities of layered situational constructs. Previous work by Dockhorn et al. (2007) addresses what could be called the context of situation development, where an "invariant" is defined as the necessary and sufficient conditions for a situation to exist. Addressing context and its importance in situation estimation is also addressed from various points of view in works by Snidaro et al. (2016). Yet other work of Costa et al. (2006) addresses a distributed rule-based approach for situation detection. When well-designed and developed, rule-based systems can be both efficient to develop and to effective to employ, but there are many lessons-learned and limitations of rule-based systems that need to be considered (e.g., Nazareth, 1989), such as scalability, blindness to data not included in the rules, and coverage of unbounded parameter values. While the foundational and systemic aspects of these works are very relevant to our discussion on situation control, the authors point out in more than one of these papers that evaluations of these prototype implementations are under development.

All of these works are focused on the estimation function and associated processes for developing capabilities to estimate situational states. Collectively, many systems engineering issues are addressed, to include data management, ontological issues, modeling of situations, and other related functions for situational estimation. In that regard, they are solid research projects but they are not directed to the holistic, closed-loop situation control process that involves decision-making once situational states are determined.

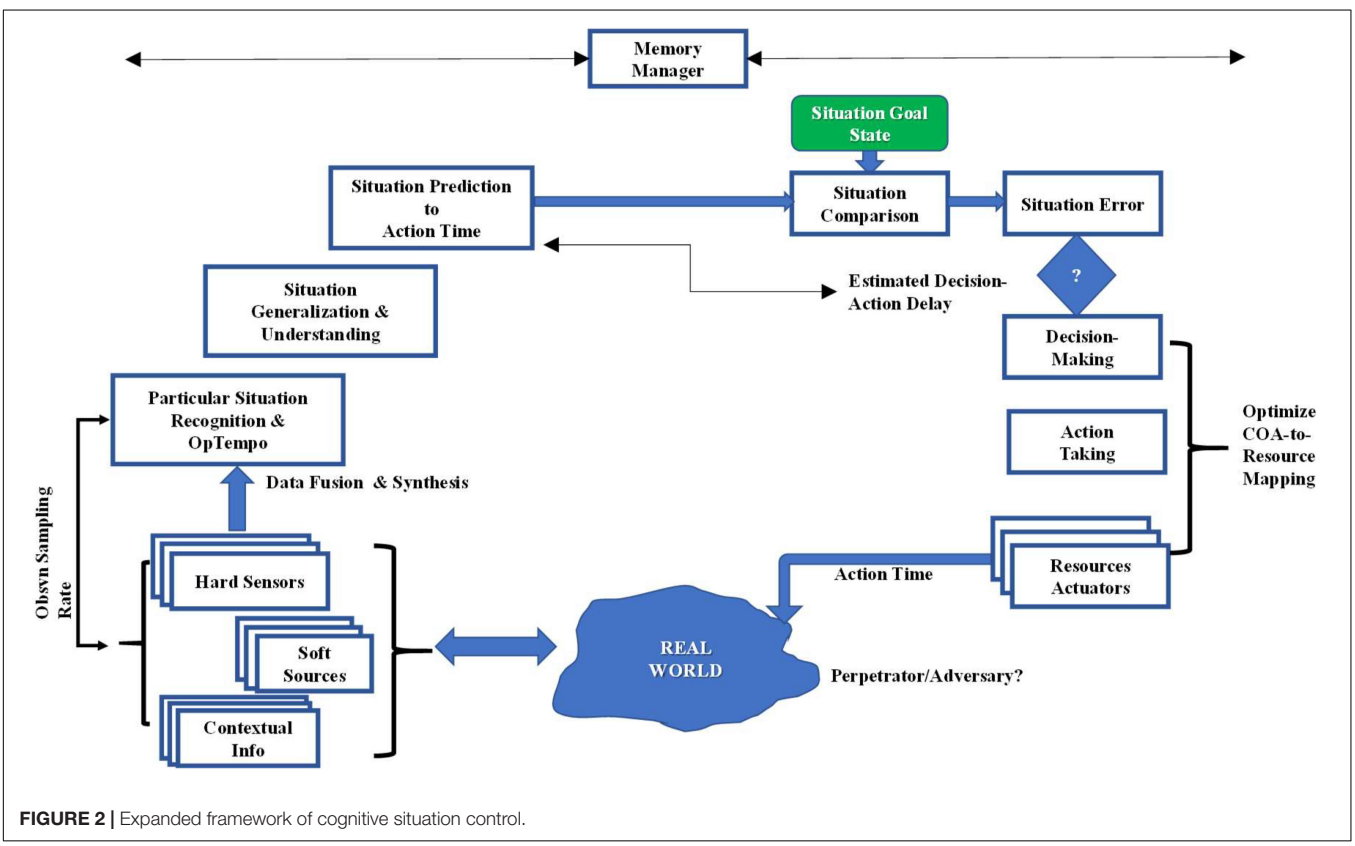


FIGURE 2 | Expanded framework of cognitive situation control.

PROPOSED FUNCTIONAL EXPANSION OF THE BASELINE FRAMEWORK: OVERVIEW

While Jakobson provides a sound initial foundation for a process description of situation control, we suggest various enhancements of this process description. A first remark is that the level of specificity of the meaning and construct of a "situation" needs to be elaborated; we see this ideally as resulting from a formal ontological development. At the highest level of abstraction, one could say that a situation is a set of entities (here, writ large, meaning not only physical entities and objects but events and behaviors) connected by a set of relations. Relations bring a new challenge to observation-based estimation because relations are not observable by conventional sensing devices, sometimes called "hard" sensors, meaning electromechanical type devices such as radars and imaging systems. Hard data produces features and attributes of entities in the situation from which inter-entity relations could be reasoned. It is possible that "soft" data such as social media data may apply to a situation control problem, in which case such data may, if based on human observation, reasoning, and judgment, yield direct estimates of inter-entity relations. Contextual type data, that imputes influences on the estimation of entities and relations, would also be fused in a robust observation and data fusion process to aid situation estimation processes. Along with the entity ontology, a relation ontology is also needed so that the specifics of a labeled, specific situational state can be assembled from these components. That assembly requires a higher level of abstraction in inferencing. Thus, Jakobson's situation recognition process will need to be supported by an ontological foundation where entities, relations, and labeled situational states are coupled to the fusion and recognition processes that will have to assemble the recognized, labeled situational state by exploiting this framework, and also by accounting for the various uncertainties in the integrated observational and inferential processes.

Another suggestion relates to the need for accounting for time. As we remarked previously, the real world is always dynamic, and so situations are in a constant process of unfolding; situations can be labeled as continuously valued random variables. Thus, we assert the need for a Situation Prediction process that is the means for maintaining situation understanding over time. How such a process may be framed depends on how the situation state is modeled; for example, a situation could be represented as a graph (entities as nodes, relations as arcs) or as a pattern of variables in the form of a time-series, or yet other representational forms. Many strategies for prediction address the problem as a pseudo-extrapolation of some type, projecting the most likely evolution of the dynamic sub-parts of a current situation. This brings in the need for Situation Understanding, which we characterize as a process that enables generalization from the particulars of the moment. Situation Understanding admits to adding knowledge and thus adding (or subtracting) new piece-parts of the situational construct, thus enabling more insightful projection of estimated situation dynamics. At some point in time or as part of an ongoing process, an assessment of whether the situation is satisfactory or not is typically carried out; this requires a specification of some desired situational state (as previously noted, Jakobson calls it a Goal Situation in **Figure 1**) that is the basis for comparison. Executing this step thus requires a process for Situational Comparison. However, executed, the comparison process yields what could be called an "error signal" as would exist in any control process, as Jakobson points out; we assert that this error signal will have stochastic properties, since the estimated situational state, and perhaps the goal state as well, will have stochastic-type error factors embedded in the calculations. The error signal requires assessment as to whether any action is required, and so there is a question as to "degree" of error, and if the error is stochastic, issues of variance in this error variable will factor into the severity assessment.

Another timing issue also arises at this point: this relates to the issue of synchronizing the action-taking and the situation prediction processes in order that the planned action is in fact acting on the intended world situation at the action-time. All these processes consume time, and an estimate of the sum of the decision-time and acting-time will set a requirement for situation prediction so that the actions that occur are acting onto the expected situation at that time; thus, there are process interdependencies (see Llinas, 2014) for further remarks on this point). These expanded remarks and functional needs are depicted in **Figure 2** that shows our suggestions for an expanded framework of situation control:

EXPANDED FUNCTIONAL REVIEW: STATE OF THE ART AND CHALLENGES

Situation Recognition

One definition of "recognize" is to "perceive something previously known," implying that a model-comparison type process is employed for recognition. But even before a model is conceptualized, a modeling framework is required to set a norm for the structure and content of such a model; this requirement brings into our discussion the need for a situation ontology. To our knowledge, no fully and well-developed, formalized ontological specification of a situational state exists that has been taken up broadly by researchers addressing the kind of problems we are discussing here (e.g., the data fusion community). There has been a fairly large number of publications that offer representational schemes for situations, some labeled as ontologically-based, but those models have not been broadly applied (see Dousson et al., 1993; Boury-Brisset, 2003; Baumgartner and Retschitzegger, 2006; Little and Rogova, 2009; Cardell-Oliver and Liu, 2010; Almeida et al., 2018, that are just a sampling). As situations are rather complex world states, processes trying to estimate these states need to take a position on what the components of situations are, as most approaches can be labeled as bottom-up, assembling situational state estimates from estimates of the components. Development of a rigorous situational ontology and harmonization of its use across a community is a very complex matter. It would seem that such an issue should fall to the portion of a community addressing its engineering methods, and the regularization of

top-down system engineering approaches; how this issue will unfold going forward remains unclear.

As it is clear that situations evolve and change over time, we need to think about the tempo of situation recognition as a process, e.g., as a "freeze-frame" depiction or perhaps an interval-based depiction. This issue muddies the distinction between recognizing a situation and prediction and updating of situations; the underlying issue is that a situational state is a continuous variable, an emphasis previously pointed out. Importantly, perhaps even crucially, the ability to assess the situation evolution rate/OpTempo is needed to specify the required observational rates of situation components, in the fashion of a "Nyquist" criterion for signal sampling. Clearly if the observational rates across the sensor suite are not tailored to the situational tempo, the entire situation estimation and control framework is compromised.

Sampling of Computational Methods for Situation Recognition

As commented above, one way that Situation Recognition (SR) can be approached is as a model-comparison process. In Dahlbom et al. (2009) a template-based approach to SR is described. This paper raises some basic questions for any model-based approach, to include deciding which situation-types to model, how complete must the matching process be, and other issues related to the model-comparison scheme. We also point out that a model-comparison approach, to include any ML approach, is based on historical, available data and *a priori* knowledge, and a root question revolves around the use of "history" to assess the "future," meaning that an argument needs to be shaped that verifies that the applicability of such models includes an acceptable spectrum of possible future situations of interest. That is, such methods have boundaries of situation coverage and will not address anything that is not modeled, such as possible effects of nature, effects of contextual factors, or the creative actions of an adversary. If we abstract a "situation" as a set of entities in a set of relations, we can say that SR aims at identifying complex constellations of entities and relations, i.e., situations, extracted from a dynamic flow of complex observational and other data and information. In a broad sense, one could say that SR is a filtering process. This process will depend on the extent and quality of both real-time data projected to be available and of the *a priori* knowledge employed in model construction. This distinction or balance of available real-time data and degree of *a priori* domain knowledge is clearly a crucial *a priori* design issue for the design of any SR approach. The requirements for either of these factors depends in part on the complexities of the set of relations embodied in any situational construct; if the relations are simple, they should be able to be inferred from observational data but if they are complex, they will need to be derived from a combination of observational data and *a priori* knowledge. In Dahlbom et al. (2009), a simplified scheme for a template-based approach to SR is developed; they point out that template-based methods have also been applied in the extensive work in plan recognition, such as in Azarewicz et al. (1989) and Carberry (2001), as well as other early AI-based techniques such as rule-based systems, both also being model-based approaches. Without doubt, the framework used most frequently for Situation Recognition is the Bayesian Network (BN)/Bayesian Belief Net (BBN) approach; the publications advocating the use of BBN are numerous. Some researchers Elsaesser and Stech (2007) suggest that BBN's "can be thought of as a graphical program script representing casual relations among various concepts represented as nodes to which observed significant events are posted as evidences," which is pretty much the dynamic process of interest here. The idea of that paper is to construct BBNs from sub–networks of internodal relations. An important advantage of this approach is that it uses BBNs distributed across multiple computers exploiting simple standard "publish" and "subscribe" functionalities that allows for significant enhancement of the inferencing efficiency. Multi-agent architectures involving other estimation techniques at the nodes are also used for Situation Recognition. Many other paradigms for SR including Fuzzy Logic and Markovian methods can be seen in the literature.

Situation Prediction

As noted in section "Proposed Functional Expansion of the Baseline Framework: Overview," the requirement for a situation prediction (SP) process is linked to the time of action onto the predicted situation. As for most prediction, projection, or extrapolation processes, the difficulty and accuracy of such processes is linked to the temporal degree of projection (how far ahead) and the rate of observation and input of any data that the projections depend on; this is not just sensor/observational data but contextual and soft data as well. We have emphasized the importance of the temporal aspects and the need to maintain situation awareness; that emphasis is acknowledged in various recent publications (e.g., Blasch, 2006; Niklasson et al., 2008; Baumgartner et al., 2010; Foo and Ng, 2013). Research areas where situation prediction has been addressed include cyber defense, for attack/intent projection, autonomous vehicles where traffic situation prediction is crucial, and also crisis/disaster management, to guide response services.

Sampling of Computational Frameworks for Situation Prediction

Two application areas where SP is addressed are those related to Cyber SP for cyber-defense and Traffic Situation SP related to autonomous car systems. A broad area where SP has also been addressed is in a wide variety of game settings, from Chess to Wargaming to Video Gaming. Most game environments, however, have various rules that can constrain the evolution of situations and thus provide a constrained framework within which to explore SP, although many other settings will also have constraints. We choose to show the SP framework of Baumgartner et al. (2010) for traffic prediction that describes a holistic approach that shows the joint exploitation of an SA Ontology and, in this case, Colored Petri Nets (CPN) as an SP estimation/modeling scheme.

In the traffic/autonomous car application, it is desired to predict critical situations from spatio-temporal relations between objects. These and other relations can be expressed by employing relation calculi, each of them focusing on a

certain spatio-temporal relation, such as mereotopology-based ("part-of" based), spatial orientation, or direction. According to Baumgartner et al. (2010) these calculi are often modeled by means of Conceptual Neighborhood Graphs (CNGs, see Freksa, 1991); as noted in this paper, the CNG's impose constraints on the existence of transitions between relations, thus providing a way to bound the complexity of relation modeling. CNGs can be used for modeling continuously varying processes, and have been used in a variety of related applications. Representing CNGs as CPNs can lead to increasing prediction precision by using precise ontological knowledge of object characteristics (if the ontology is done well) and interdependencies between spatio-temporal relations. This can lead to increased prediction explicitness in their approach by associating transitions with dynamically derived distances for multiple view-points. These so-called Situation Prediction Nets (SPN) in Baumgartner et al. (2010) are derived automatically from the available situation awareness ontologies. The research described in this paper is among the few that proactively integrate an ontological framework for relations and situation structures with a computational strategy for SP.

In Salfinger et al. (2013), a situation's evolution is modeled as a sequence of object-relational states it has evolved through, i.e., the sequence of its situation states. This approach discretizes the continuous evolution of the monitored real-world objects into a sequence of their different joint relational states defined by various relations between those objects, defined in an "alphabet" or what could be called a bounded ontology. Thus, the problem of predicting a monitored situation's evolution is cast as a sequence prediction problem. This technique is also applied here to the traffic-situation prediction problem. This approach employs a Discrete Time Markov Chain scheme; this is preceded by a situation-mining analysis to define the situation state-space "alphabet," learned from human-labeled state sequences.

As previously remarked, works on SP can also be found in the cyber-defense domain. In Husák et al. (2019), a survey of such methods is provided. Their approach addresses four categories of predicative capability. The first two of these categories are attack projection and intention recognition, in which there is a need to predict the next move or the intentions of the attacker, third is intrusion prediction, in which predictions are developed of upcoming cyber-attacks, and fourth is network security situation forecasting, in which projections are made of the cybersecurity situation in the whole network. Across these applications, the paper reviews two broad categories of prediction techniques: discrete-time approaches, and continuous-time approaches. The discrete-time techniques include: "attack graphs" that probabilistically model initial and successor states of a postulated attack process. As in Salfinger et al. (2013), the state-space if often defined by a data mining analysis. The predictions using attack graphs are based on traversing the graph from an initial state and searching for a successful or most-probable attack path. A number of papers are cited in the survey that employ variations of this technique. Bayesian Nets and Markov techniques, as well as Game-theoretic methods are among the other discrete-time approaches reviewed. The continuous-time methods reviewed fell in to two categories, time-series methods and "gray" methods. These methods were largely used for whole-network predictions involving forecasts of the numbers, volumes, and composition of attacks in the network and their distribution in time.

Some Views From Cognitive Neuroscience

We have maintained that Situation Prediction is a functional requirement in the process of Situation Control. There are relatively few frameworks offered in the technical engineering literature for Situation Prediction (as just discussed) but there are also some paradigms for this process in the computational neuroscience literature. For example, Bubic et al. (2010) provide one overview of such processes in the brain. In this paper, distinctions are made in relation to the horizon over which predictions might be made (as we have also mentioned previously). For example, the term "expectation" is said to reflect the information regarding the spatial and temporal characteristics of an expected event, whereas "anticipation" describes the impact of predictions on current behavior, e.g., decisions and actions based on such predictions, and "prospection" is described as an ability to "pre-experience the future by simulating it in our minds." These distinctions are shown in **Figure 3**.

The main factors that influence the nature of a predictive process are characterized in Bubic et al. (2010) as shown in **Figure 4**.

Heeger, in a paper that provides somewhat detailed mathematical models of cortical processes (Heeger, 2017), suggests that prediction is one of three key cortical operations: (i) inference: where perception is a non-convex optimization that combines sensory input with prior expectation; (ii) exploration: here, inference relies on neural response variability to explore different possible interpretations; and (iii) prediction: inference includes making predictions over a hierarchy of timescales, not unlike suggested by Bubic et al. (2010) The starting point for this development is the hypothesis that neural responses minimize an energy function that represents a compromise between the feedforward drive and prior drive (drive $\approx$ neural signals). In these process models, the responses of the full population of neurons (across all channels and all layers) are asserted to converge to minimize a global optimization criterion, which Heeger calls an energy function. Specifically, the starting point for this model development is the hypothesis that neural responses minimize an energy function that represents a compromise between the feedforward drive and prior drive. Heeger says that predictive coding theories start with a generative model that describes how characteristics of the environment produce sensory inputs; Perception on the other hand is presumed to perform the inverse mapping, from sensory inputs to characteristics of the environment. Heeger's approach suggests a different process for how the brain might predict over time, relying on a recursive computation similar to a Kalman filter, where the predictive basis functions serve the same role as the dynamical system model in a Kalman filter.

Returning to cognition, many researchers in the neuro and cognitive sciences have developed a view according to which prediction or anticipation represents a fundamental characteristic of brain functioning, suggesting that prediction

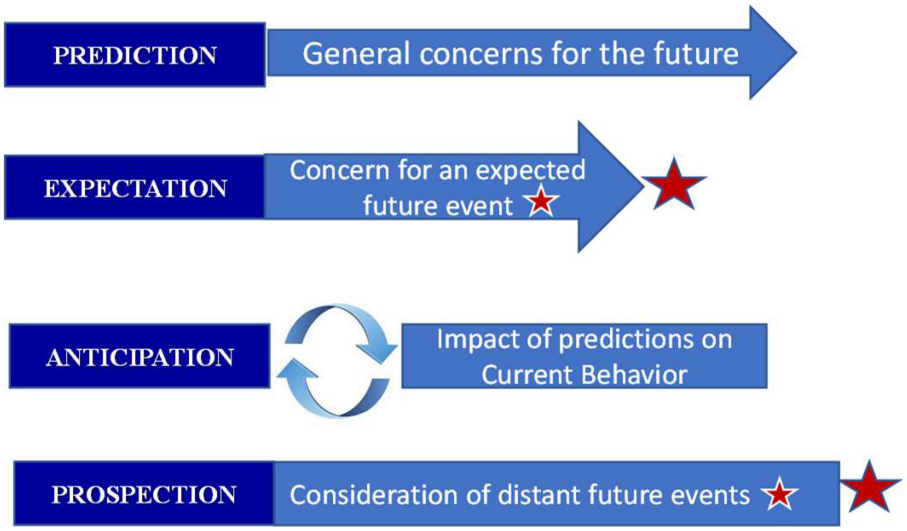


FIGURE 3 | Distinctions in prediction-anticipation-prospection derived from Bubic et al. (2010).

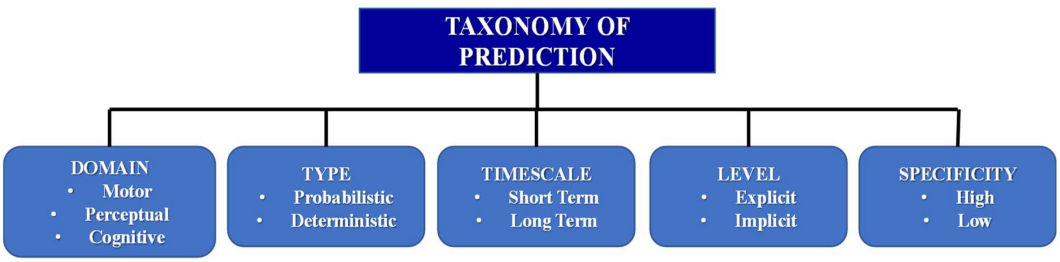


FIGURE 4 | Factors influencing the nature of prediction derived from Bubic et al. (2010).

is "at the core of cognition" (Pezzulo et al., 2007). Further, for many cognitive functions and neural systems, an ability to anticipate is a core requirement, such as in motor and visual processing and attention (Mehta and Schaal, 2002). According to Friston (2005), predictive processing is inherent to all levels of our organized neural system. It is suggested that predictions drive our perception, cognition, and behavior in trying to fulfill predictions by preferentially sampling features in the environment. Nevertheless, it can be expected that mismatches will occur, and the size of such mismatches (prediction error) creates a "surprise" that the brain tries to minimize in order to maintain present and future stability (Friston and Stephan, 2007). In reviewing Bubic's paper, one comes away with the interpretation that anticipatory or predictive processing potentially reflects one of the core, fundamental principles of brain functioning which justifies the notion of "the predictive brain" seen in some papers.

These neuronal-level models are quite interesting in helping to understand how the brain develops predictions, but what is being predicted are anticipated human-based sensor signals that are important to human survival.

One such model, the Virtual Associative Network (VAN), is combined with active inference and presented elsewhere in this Frontiers special edition (Moran et al., 2021). This work presents a new, Cognitive-Partially Observable Markov Decision Process (C-POMDP) framework, extending the Partially Observable Markov Decision Process (POMDP) to account for an internal, cognitive model which attempts to contend with situation control considerations we outline here such as situation recognition, prediction, learning and understanding.

The C-POMDP framework presumes an active interaction between the agent and its environment wherein the agent interacts with the environment in repetitive cycles consisting of (i) sensing observable phenomena within the environment; (ii) estimating situational states, situation dynamics (behavior, op tempo, relations, etc.); (iii) predicting future states and rewards; and (iv) making decisions to maximize expected rewards. A key point here is that these estimation and decision-making processes are based upon an internal model which is maintained and updated by the agent as it reasons about experiences. In Moran et al. (2021), learning is facilitated by probabilistic reasoning and operations on a graph-based modeling structure which encapsulates associations between objects (entities, situation artifacts), behaviors, and relations.

Situation Learning and Situation Understanding

The topics of learning and understanding have of course been extensively studied by a variety of research and application communities. These concepts have some relationship but they are also distinct from each other. Learning can be seen as dependent on (at least) two processes: observation and data gathering, and on experimentation and acting. Both processes produce real-time data that support inductive processes directed to gaining real-time knowledge. Understanding would seem to follow learning wherein the gained knowledge, along with archived knowledge, are exploited *in combination* to develop a *generalized* understanding of a world situation that allow development of a contextual perspective—a generalized perspective—of that situation. Generalization allows the recognition of the similarities in knowledge acquired in one circumstance, allowing for transfer of knowledge onto new situations. The knowledge to be transferred is often referred to as abstractions, because the learner abstracts a rule or pattern of characteristics from previous experiences with similar stimuli. Yufik (2018) defines understanding as a form of active inference in self-adaptive systems seeking to expand their inference domains while minimizing metabolic costs incurred in the expansions; the process thus also entails an optimization element directed at minimizing neuronal energy consumption. This view also sees understanding as an advanced adaptive mechanism in virtual associative networks involving self-directed construction of mental models establishing relations between domain entities. Understanding inter-entity relations is also a core element of situation understanding. Thus, the relationship between learning and understanding can be seen as complementary; understanding complements learning and serves to "overcome the inertia of learned behavior" when conditions are unfamiliar or deviate from those experienced in the past (Yufik, 2018). A challenge now receiving considerable attention with the new thrusts into AI is to understand how humans are able to generalize from very limited sampling. One approach fostered by Tenenbaum et al. (2011) and Lake et al. (2015) is based on probabilistic generative models, proposed as a basis for linking the psychological and physical aspects of the world. These techniques are being explored in DARPA's Machine Commonsense program; however, these techniques will yield learning and understanding processes that create the foundational nuggets of what humans typically call "common sense" knowledge, often called tacit knowledge, and are far from a computational ability to understand situations of varying complexity. (An often-cited example of common, tacit knowledge that humans accrue is the learning of embedded rules of grammar that are learned over time from discrete sampling.) As most would agree that understanding involves uncertainty, whereas knowledge is often defined as "justified true belief" following Plato (yet acknowledging Gettier),[3] it seems reasonable to explore probabilistic methods to model commonsense understanding. The issue of exactly how certain one must be about a belief to qualify as "knowing" has been called the boundary problem (Quine, 1987). We see that there are thus distinctions between understanding and knowledge; importantly, understanding can be possibly incorrect. Also important to this discussion, as just mentioned, is the process of generalization, a rather pervasive topic in psychology. In Austerweil et al. (2019), discuss the issue of learning how to generalize, which suggests that generalization requires postulating "overhypotheses" or constraints in effect on the hypothesis domain to be nominated. Some assert that such overhypotheses are innate but Austerweil et al. (2019) argue that they can be learned. In either case, the generalization framework is said to be Bayesian-based. Generalization has also been studied in Shepard (1987) that suggests an exponential metric distance between the stimuli as a basis to assert similarity, and in Kemp et al. (2006) that discusses the overhypotheses issue. We note that the issue of assessing similarity or degrees of association between disparate or multimodal data is broadly similar to the generalization question, and is a topic addressed in the field of multisensor data fusion. In those cases, techniques of multidimensional scaling, copulas, and manifolds have been used to develop scaling methods to relate such non-commensurate data.

Situation Comparison

The assessment of any situation as to its acceptability or to the need for situation control and action-taking requires the specification of some basis for comparison; in **Figure 1** Jakobson shows the Situation Comparator function needing a Goal Situation to be defined. As situational states can be rather complex, the bases of comparison could perhaps be done for portions of a situation rather than the entirety of a complicated, entangled set of situational elements. How any such comparisons would be done is also dependent on how one chooses to represent situations. Our search for literature related to this situation comparison issue shows that this issue has not been extensively addressed, and the methods proposed are of very different type, as described next.

Sampling of Computational Methods for Situation Comparison

In Mannila and Ronkainen (1997), as in other works reviewed here, a situation is depicted as a series of events, i.e., an event sequence. For many papers, as we will see, the issue of comparison evolves around developing notions of similarity. In Mannila and Ronkainen (1997) then, there is the issue of defining similarity across event sequences. Building on the intuition that differences or similarities in sequences relates to how much work has to be done to convert one sequence to another, they define an "edit distance" measure of similarity. These edit distance measures are computed using a dynamic programming approach. Sequence transformation operations such as insert, delete, and move are formed, as well as a cost measure. From this framework, an optimization function can be developed to compute the minimum cost of a sequence edit between sequence pairs. Some limited empirical results are developed that show reasonable performance of this exploratory approach. Sidenbladh et al. (2005) propose an approach based on using random sets as the representational form for situations. This paper compares rolling

[3]https://en.wikipedia.org/wiki/Gettier_problem

situation predictions as a use case where the situation predictions at two different times are normalized due to estimation noise differences, arguing that prediction error is proportional to prediction time. Given that normalization, they define a standard norm as a similarity/difference measure and also point out that the Kullback-Liebler measure[4] is inappropriate for this purpose. In some of our own work, we have depicted situations as graphs, following a simple situation definition, as previously mentioned, as a set of entities connected by a set of relations. Situation similarity then can be assessed by any of the many existing types of metrics for graph comparison (see e.g., Hernandez and Van Mieghem, 2011). Which metrics are best will depend on the graph details; for example, relations among entities can be directed, and so comparison would then require metrics that account for directed arcs in the representational graphs for the situations being compared. There are metrics that can compare both the global and local characteristics of two graphs; methods of this type have been used for anomaly detection in situational analysis. Since the description of any situational state will employ language to label the situational components (entities) and their relations, notions of situational similarity may also involve issues of semantic similarity in the terms employed. Our research in hard and soft data fusion for disaster response needed to address this issue, which has been studied extensively, since semantic similarity and whether words mean the same thing is a core issue in many application settings. A hierarchically structured ontology or taxonomy can be useful in estimating the semantic similarity between nodes in the taxonomic network. Two specific approaches used to determine the conceptual similarity of two terms in this type of network are known as node and edge-based approaches. The node-based approach relates to the information content approach while the edge-based approach corresponds to the conceptual distance approach. The edge-counting measures are based on a simplified version of spreading activation theory (Cohen and Kjeldsen, 1987) that asserts that the hierarchy of concepts in an ontology is organized along the lines of semantic similarity. Thus, the more similar two concepts are, the more links there are between the concepts, and the more closely related they are Rada et al. (1989). The node-based measures are based on the argument that the more information two terms share in common, the more similar they are, and the information shared by two terms is indicated by the information content of the terms that subsume them in the taxonomy. Data association methods employed in data fusion have been used to assess whether two situation states have the same objects in them (e.g., Stubberud and Kramer, 2005); these metrics used ideas from metric spaces and cardinality principles to compute object-set similarities. Other techniques for assessing situational similarity can be drawn from measures for assessing similarity of sets such as the Jaccard Similarity and the Overlap Coefficient (Rees, 2019).[5] Similarity of relations is also of interest, and the methods of ontological similarity could be used for relation-labels as well as methods from Fuzzy Logic and latent variable type analyses (Turney, 2006). Finally, (Gorodetsky et al., 2005) develop a situation updating method that addresses the issue of asynchronous data with a data ageing scheme, the missing data issue with a direct data mining approach, and a situational state classification scheme based on a rule-based approach, in an effort to account for these various aspects of situation updating in an integrated approach.

CONTROL DYNAMICS

We have described the overall control process so far as rather linear and feed-forward but there may be inter-functional interdependencies across each "situational" function described here. As multisensor data fusion processes are relevant information processes supporting situation control as candidate processes for situation estimation (see, e.g., Liggins et al., 2009) it is known that there can be inter-process dependencies that need to be addressed among data fusion, situation estimation, and decision-making processes (see Llinas et al., 2004; Llinas, 2014). In the case of data fusion processes, the approach to situation estimation is typically layered, following a "divide and conquer" approach typically employed for complex problems. The layered estimates are partitioned according to specificity, with lower levels estimating features of situational entities, and upper levels estimating aggregated multi-entity relational constructs. Thus, the layers share content about common entities that may be helpful to share in a synergistic scheme; for data fusion, Llinas et al. (2004) addresses some of the issues of this point. In the case where data fusion and decision-making processes are integrated in a single architecture, the inter-process dependencies exist because one process, data fusion, is estimating a situation and the other process is deciding about situations; these interdependencies are discussed in Llinas (2014). Further, the Action Planning and Action-Taking processes that depend on the possibly complex viable action-spaces of available resources (that is, the various situation-affecting actions that a given resource can execute) can lead to the need for an optimization-based approach to select the best resource to execute a particular situation-affecting action. Situation OpTempo and overall timing control again need to be considered since there can be delays in making the action-taking decisions (e.g., solving an optimization problem) and delays in employing a resource and realizing its intended effects. Consideration of these factors aids in estimating the time it takes to make a decision and the time for resources to act on the situation. An *a priori*/ongoing estimate of the sum of these times provides the time specification to the Situation Prediction function so that the system is predicting the situational state at the expected time of action from the resources; also discussed in Llinas (2014).

Partially Observable Markov Decision Process

Control Theory offers a foundational problem formulation for many problems requiring Situation Control. Such problems presume an active interaction between an intelligent agent and its environment where:

[4]https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

[5]https://medium.com/rapids-ai/similarity-in-graphs-jaccard-vs.-the-overlap-coefficient-610e083b877d

- The agent exercises repetitive cycles of sensing the environment, executing actions and modifying them based upon feedback
- The agent seeks to maximize cumulative rewards received from the environment
- The agent iteratively maps an error signal into actions.

In the POMDP formulation, these problem elements are expressed as sets, and mappings between the sets. More specifically, the environment offers a set of states (S) and a set of rewards (R). The agent will iteratively draw from a set of observations (O), and choose from a set of actions (A). Here, the dynamics of the situation are characterized by a set of state transition probabilities (P), providing a mapping from a particular state at time t to a state at time $t + 1$ (P: $s_t \rightarrow s_{t+1}$). The agent's observations which are related to environmental state (S), are characterized by a set of observation probabilities (Z) which map state at time t to an observation at a time t or a later time $t + n$ (Z: $s_t \rightarrow o_{t+n}$, $n = 0$). Similarly, the relationship between rewards and underlying state received by the agent may be modeled deterministically or stochastically as related to state; If stochastic, the relationship between the states (S) and rewards (R) will be characterized by a reward probability mapping, Q (Q: $s_t \rightarrow r_{t+m}$, $m \geq 0$). For further information on a POMDP modeling approach, the reader is referred to Bertsekas (1987).

Although the POMDP offers a principled problem formulation for complex situation control problems, it is well established that, for realistic problems, POMDP solutions often suffer from "the curse of combinatorial explosion" and approximate solutions methods are required for solution (Bertsekas, 1987). These approximate methods include, perhaps most notably, Reinforcement Learning methods (Spaan, 2012) which have been used extensively in some artificial intelligence solutions.

The authors contend that the POMDP offers a starting point for the control aspects of the situation control problem formulation but effective solutions for complex situation control problems will require that the relationships between pertinent situational factors governing state transition probabilities (P), observation probabilities (Z), and reward probabilities (Q) be understood. In practice, identifying the relevant situational factors and accurately modeling the relationships governing these mappings will be derived experientially, through situation learning as described in section "Situation Learning and Situation Understanding" above. Further, the temporal considerations we have cited such as the situation's *op tempo* guiding the agent's observation rate, and the need for situation prediction over multiple horizons accounting for both state and action dynamics, must be taken into account in order to properly assess *situation error*, a key step in the process model.

SUMMARY

There is a large literature on Situation Awareness and Situation Assessment that, to a large degree, treats the estimation of these states in isolation from many other functions needed to frame a complete, closed-loop process that not only estimates these states but addresses the overarching central issue for so many applications of situation control. Jakobson and a number of others, largely from the community of authors and attendees of the IEEE Cognitive Situation Management (CogSIMA) Conferences, have addressed many issues related to situation control and have tried to move the science forward by expanding the process view to a more holistic framework. This paper is a contribution to that collection of works, and also offers some limited remarks from the point of view of computational neurodynamics that is intended to lay the foundation for a dialog regarding the exploitation of Machine Intelligence within and central to the situation control paradigm. This is a complex space of thinking, of process architecting, of algorithmic design and development, and of human-machine interaction. As the broad technical communities of the world grapple with the development and exploitation of AI, ML, Machine Intelligence, and of the role of humans and of autonomous systems and behaviors, the need to frame the situation control process will be a central topic in the broadest sense; this paper is a small contribution to that goal.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

The manuscript was invited for a special addition entitled "Understanding in the Human and the Machine" and fees will be paid for as arranged.

ACKNOWLEDGMENTS

JL would like to acknowledge the support of the Sensors Directorate of the U.S. Air Force Research Laboratory for providing support for this research.

REFERENCES

Almeida, J. P. A., Costa, P. D., Guizzardi, G., and João, P. A. (2018). "Towards an ontology of scenes and situations," in *Proceedings of the 2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. (Boston, MA), 29–35. doi: 10.1109/COGSIMA.2018.8423994

Austerweil, J. L., Sanborn, S., and Griffiths, T. L. (2019). Learning how to generalize. *Cogn. Sci.* 43:e12777. doi: 10.1111/cogs.12777

Azarewicz, J., Fala, G., and Heithecker, C. (1989). "Template-based multi-agent plan recognition for tactical situation assessment," in *In: Proceedings of the Fifth IEEE Conference on Artificial Intelligence Applications*. (Miami, FL: IEEE).

Baumgartner, N., Gottesheim, W., Retschitzegger, W., Mitsch, S., and Schwinger, W. (2010). "Situation prediction nets – playing the token game for ontology-driven situation awareness," in *Proceedings of 29th International Conference on Conceptual Modeling*. (Vancouver, BC: Springer). doi: 10.1007/978-3-642-16373-9_15

Baumgartner, N., and Retschitzegger, W. (2006). "A survey of upper ontologies for situation awareness," in *Proc. of the 4th Intl. Conf. on Knowledge Sharing and Collaborative Engineering*. (Calgary,AB: ACTA Press), 1–9.

Bertsekas, D. P. (1987). *Dynamic Programming, Deterministic and Stochastic Models*. New Jersey, NJ: Englewood Cliffs, 07632.

Blasch, E. (2006). Issues and challenges in situation assessment (Level 2 fusion). *Artif. Intell.* 1, 122–139. doi: 10.21236/ADA520878

Boury-Brisset, A.-C. (2003). "Ontology-based approach for information fusion," in *Proceedings of the Sixth International Conference on Information Fusion*. (Cairns,AU-QLD: IEEE), 522–529. doi: 10.1109/ICIF.2003.177491

Boyd, J. R. (1986). Patterns of Conflict. (Unpublished briefing).

Bruner, J. S. (1973). *Beyond the Information Given*. New York, NY: W.W. Norton and Company.

Bubic, A., von Cramon, D. Y., and Schubotz, R. I. (2010). Prediction, cognition and the brain. *Front. Hum. Neurosci.* 4:25. doi: 10.3389/fnhum.2010.00025

Carberry, S. (2001). Techniques for plan recognition. *User Model. User-Adapt Interact.* 11, 31–48. doi: 10.1023/A:1011118925938

Cardell-Oliver, R., and Liu, W. (2010). Representation and recognition of situations in sensor networks. *IEEE Commun. Mag.* 48, 112–117. doi: 10.1109/MCOM.2010.5434382

Cohen, P. R., and Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manag.* 23, 255–268. doi: 10.1016/0306-4573(87)90017-3

Costa, D., Guizzardi, G., Almeida, J. P. A., Ferreira Pires, L., and van Sinderen, M. (2006). "Situations in conceptual modeling of context," in *In Proceedings of the Workshop on Vocabularies, Ontologies, and Rules for the Enterprise (VORTE 2006) at IEEE EDOC 2006*. (Hong Kong: IEEE Computer Society Press). doi: 10.1109/EDOCW.2006.62

Costa, P. D., Mielke, I. T., Pereira, I., and Almeida, J. P. A. (2012). "A model-driven approach to situations: Situation modeling and rule-based situation detection," in *in Proceedings of EDOC. IEEE*. (Beijing: IEE), 154–163. doi: 10.1109/EDOC.2012.26

Dahlbom, A., Niklasson, L., Falkman, G., and Loutfi, A. (2009). "Towards template-based situation recognition," in *Proceedings of SPIE Defense, Security, and Sensing*, Vol. 7352a. (Orlando, FL: SPIE). doi: 10.1117/12.818715

Dockhorn, P. C., Almeida, P. A., Pires, L. F., and van Sinderen, M. (2007). "Situation specification and realization in rule-based context-aware applications," in *In Proc. of the Int. Conference DAIS'07*, Vol. 4531. (Paphos: Springer), 3247.

Dousson, C., Gaborit, P., and Ghallab, M. (1993). "Situation recognition: representation and algorithms," in *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*. (Chambery), 166–172.

Elsaesser, C., and Stech, F. (2007). "Detecting deception, chap 2.1," in *Adversarial Reasoning: Computational Approaches to Reading the Opponent's Mind*, eds A. Kott and W. McEneaney (Boca Raton, FL: CRC Press). doi: 10.1201/9781420011012.ch2.1

Foo, P. H., and Ng, G. W. (2013). High-level information fusion: an overview. *J. Adv. Inf. Fusion* 8, 33–72.

Freksa, C. (1991). "Conceptual neighborhood and its role in temporal and spatial reasoning," in *In Proc. of the Imacs Intl. Workshop on Decision Support Systems and Qualitative Reasoning, pages*. (Munich), 181–187.

Friston, K. J. (2005). A theory of cortical responses. philos. *Trans. R. Soc. Lond., B, Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Gorodetsky, V., Karsaev, O., and Samoilov, V. (2005). On-line update of situation assessment: a generic approach. *Int. J. Knowl. Based Intell. Eng. Syst.* 9, 351–365. doi: 10.3233/KES-2005-9410

Heeger, D. J. (2017). Theory of cortical function. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1773–1782. doi: 10.1073/pnas.1619788114

Hernandez, J. M., and Van Mieghem, P. (2011). *Classification of graph metrics Tech. Rep 20111111*. Delft: Delft University of Technology.

Hristidis, V., Chen, S. C., Li, T., Luis, S., and Deng, Y. (2010). Survey of data management and analysis in disaster situations. *J. Syst. Softw.* 83, 1701–1714. doi: 10.1016/j.jss.2010.04.065

Husák, M., Komárková, J., Bou-Harb, E., and Celeda, P. (2019). Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Commun. Surv. Tuts* 21, 640–660. doi: 10.1109/COMST.2018.2871866

Jakobson, G. (2008). "Introduction to cognitive situation management for tactical operations," in *IEEE Communications Society Distinguished Lecture*. (Goteborg).

Jakobson, G., Buford, J., and Lewis, L. (2006). "A framework of cognitive situation modelling and recognition," in *Proceedings of the 25th IEEE Military Communications Conference (MILCOM)*. (Washington, DC). doi: 10.1109/MILCOM.2006.302076

Jakobson, G., Buford, J., and Lewis, L. (2007). "Situation Management: Basic Concepts and Approaches," in *in Proceedings of the 3rd International Workshop on Information Fusion and Geographic Information Systems*. (St. Petersburg).

Jakobson, G. A. (2017). "framework for cognitive situation control. in Cognitive and Computational Aspects of Situation Management (CogSIMA)," in *in Proceedings of 2017 IEEE Conference on Cognitive Situation Management*. (Savannah, GA). doi: 10.1109/COGSIMA.2017.7929577

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2006). "Learning overhypotheses," in *In Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. (Hillsdale, NJ), 417–422.

Klein, G., Phillips, J. K., Rall, E. L., and Peluso, D. A. (2007). "A data-frame theory of sensemaking," in *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, ed. R. R. Hoffman (Hove: Psychology Press), 113–155.

Kokar, M. M., Matheus, C. J., and Baclawski, K. (2009). Ontology-based situation awareness. *Inf. Fusion* 10, 83–98. doi: 10.1016/j.inffus.2007.01.004

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human level concept learning through probabilistic program induction. *Science* 350, 1332-1338. doi: 10.1126/science.aab3050

Liggins, M. E., Hall, D. L., and Llinas, J. (2009). *Handbook of Multisensor Data Fusion: Theory and Practice*, 2nd Edn. Boca Raton, FL: CRC Press.

Little, E. G., and Rogova, G. L. (2009). Designing ontologies for higher level fusion. *Inf. Fusion* 10, 70–82. doi: 10.1016/j.inffus.2008.05.006

Llinas, J. (2014). "Reexamining information fusion-decision making inter-dependencies, in cognitive methods in situation awareness and decision support (CogSIMA)," in *Proceedings of the IEEE International Inter-Disciplinary Conference on Cognitive Situation Management*. (San Antonio, TX), 1–6. doi: 10.1109/CogSIMA.2014.6816532

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., and White, F. (2004). Revisions and extensions to the jdl data fusion model II," in *Proceedings of the 7th International Conference on Information Fusion*. (Stockholm).

Mannila, H., and Ronkainen, P. (1997). "Similarity of event sequences," in *In Proceedings of TIME '97: 4th International Workshop on Temporal Representation and Reasoning*. (Dayton Beach, FL), 136–139. doi: 10.1109/TIME.1997.600793

Mehta, B., and Schaal, S. (2002). Forward models in visuomotor control. *J. Neurophysiol.* 88, 942–953. doi: 10.1152/jn.2002.88.2.942

Moran, R., Friston, K., and Yufik, Y. (2021). *Active-Inference based Artificial Intelligence for Satellite Communications: A worked example of Machine Understanding*.

Nazareth, D. L. (1989). Issues in the verification of knowledge in rule-based systems. *Int. J. Man Mach. Stud.* 30, 255–271. doi: 10.1016/S0020-7373(89)80002-1

Niklasson, L., Riveiro, M., Johansson, F., Dahlbom, A., Falkman, G., Ziemke, T., et al. (2008). "Extending the scope of situation analysis," in *in Proceedings of 11th Intl. Conf. on Information Fusion*. (Cologne).

Pezzulo, G., Hoffmann, J., and Falcone, R. (2007). Anticipation and anticipatory behavior. *Cogn. Process.* 8, 67–70. doi: 10.1007/s10339-007-0173-z

Pirolli, P., and Card, S. (2005). "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *In Proceedings of International Conference on Intelligence Analysis*. (Virginia: McLean), 6.

Quine, W. V. (1987). *Quiddities: An Intermittently Philosophical Dictionary*. Cambridge, MS: Harvard Univ Press. doi: 10.4159/9780674042438

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern. Syst.* 19, 17–30. doi: 10.1109/21.24528

Richards, C. (2012). *Boyd's OODA Loop*. Atlanta, GA: J. Addams & Partners.

Rees, B. (2019). *Similarity in graphs: Jaccard versus the Overlap Coefficient*. Available online at: https://medium.com/rapids-ai/similarity-in-graphs-jaccard-versus-the-overlap-coefficient-610e083b877d (accessed November 20, 2021).

Salfinger, A., Retschitzegger, W., and Schwinger, W. (2013). "Maintaining situation awareness over time – a survey on the evolution support of situation awareness systems," in *2013 Conference on Technologies and Applications of Artificial Intelligence*. (Taipei), 274–281. doi: 10.1109/TAAI.2013.62

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243

Sidenbladh, H., Svenson, P., and Schubert, J. (2005). "Comparing future situation pictures," in *Proceedings of the 8th International Conference on Information Fusion*, Philadelphia, PA. doi: 10.1109/ICIF.2005.1591962

Snidaro, L., Garcia-Herrera, J., Llinas, J., and Blasch, E. (2016). *Context Enhanced Information Fusion*. Berlin: Springer. doi: 10.1007/978-3-319-28971-7

Spaan, M. (2012). "Partially observable markov decision processes," in *Reinforcement Learning: State of the Art*, eds M. A. Wiering and M. van Otterlo (Berlin: Springer), 387–414. doi: 10.1007/978-3-642-27645-3_12

Stubberud, S., and Kramer, K. A. (2005). "Incorporation of uncertainty into level 2 fusion association metrics," in *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP Conference*. (Melbourne, VIC). doi: 10.1109/ISSNIP.2005.1595580

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279-1285. doi: 10.1126/science.1192788

Turney, P. D. (2006). Similarity of semantic relations. *Comput. Linguist.* 32, 379–416. doi: 10.1162/coli.2006.32.3.379

Van de Walle, B., Brugghemans, B., and Comes, T. (2016). Improving situation awareness in crisis response teams: an experimental analysis of enriched information and centralized coordination. *Int. J. Hum. Comput. Stud.* 95, 66–79. doi: 10.1016/j.ijhcs.2016.05.001

Yufik, Y. (2018). "GNOSTRON: a framework for human-like machine understanding," in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence*. (Bangalore), 136–145. doi: 10.1109/SSCI.2018.8628650

Zambrano, O. M., Zambrano, A. M., Esteve, M., and Palau, C. (2017). An innovative and economic management of earthquakes: early warnings and situational awareness in real time. *Wirel. Pub. Saf. Netw.* 3, 19–38. doi: 10.1016/B978-1-78548-053-9.50002-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Systems Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 30 September 2022
DOI 10.3389/fnsys.2022.787659

OPEN ACCESS

EDITED BY
Rosalyn J. Moran,
King's College London,
United Kingdom

REVIEWED BY
William De Cothi,
University College London,
United Kingdom
Akira Taniguchi,
Ritsumeikan University, Japan

*CORRESPONDENCE
Adam Safron
asafron@gmail.com

RECEIVED 01 October 2021
ACCEPTED 02 September 2022
PUBLISHED 30 September 2022

CITATION
Safron A, Çatal O and Verbelen T (2022) Generalized Simultaneous Localization and Mapping (G-SLAM) as unification framework for natural and artificial intelligences: towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition. Front. Syst. Neurosci. 16:787659. doi: 10.3389/fnsys.2022.787659

Generalized Simultaneous Localization and Mapping (G-SLAM) as unification framework for natural and artificial intelligences: towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition

Adam Safron[1,2,3]*, Ozan Çatal[4] and Tim Verbelen[4]

[1]Center for Psychedelic and Consciousness Research, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Cognitive Science Program, Indiana University, Bloomington, IN, United States, [3]Institute for Advanced Consciousness Studies, Santa Monica, CA, United States, [4]IDLab, Department of Information Technology, Ghent University—imec, Ghent, Belgium

Simultaneous localization and mapping (SLAM) represents a fundamental problem for autonomous embodied systems, for which the hippocampal/entorhinal system (H/E-S) has been optimized over the course of evolution. We have developed a biologically-inspired SLAM architecture based on latent variable generative modeling within the Free Energy Principle and Active Inference (FEP-AI) framework, which affords flexible navigation and planning in mobile robots. We have primarily focused on attempting to reverse engineer H/E-S "design" properties, but here we consider ways in which SLAM principles from robotics may help us better understand nervous systems and emergent minds. After reviewing LatentSLAM and notable features of this control architecture, we consider how the H/E-S may realize these functional properties not only for physical navigation, but also with respect to high-level cognition understood as generalized simultaneous localization and mapping (G-SLAM). We focus on loop-closure, graph-relaxation, and node duplication as particularly impactful architectural features, suggesting these computational phenomena may contribute to understanding cognitive insight (as proto-causal-inference), accommodation (as integration into existing schemas), and assimilation (as category formation). All these operations can similarly be describable in terms of structure/category learning on multiple levels of abstraction. However, here we adopt an ecological rationality perspective, framing H/E-S functions as orchestrating SLAM processes within both concrete and abstract hypothesis spaces. In this navigation/search process, adaptive cognitive equilibration between assimilation and accommodation involves balancing

tradeoffs between exploration and exploitation; this dynamic equilibrium may be near optimally realized in FEP-AI, wherein control systems governed by expected free energy objective functions naturally balance model simplicity and accuracy. With respect to structure learning, such a balance would involve constructing models and categories that are neither too inclusive nor exclusive. We propose these (generalized) SLAM phenomena may represent some of the most impactful sources of variation in cognition both within and between individuals, suggesting that modulators of H/E-S functioning may potentially illuminate their adaptive significances as fundamental cybernetic control parameters. Finally, we discuss how understanding H/E-S contributions to G-SLAM may provide a unifying framework for high-level cognition and its potential realization in artificial intelligences.

KEYWORDS

SLAM, free energy principle, active inference, hippocampal and entorhinal systems, hierarchical generative models, robotics, artificial intelligence

Introduction

> "*We take almost all the decisive steps in our lives as a result of slight inner adjustments of which we are barely conscious.*"
> —*W.G. Sebald.*

> "*Not all those who wander are lost.*"
> —*J.R.R. Tolkien, The Riddle of Strider, The Fellowship of the Ring.*

> "*We shall not cease from exploration*
> *And the end of all our exploring*
> *Will be to arrive where we started*
> *And know the place for the first time.*"
> —*T.S. Elliot, Little Gidding.*

Autonomous systems face a fundamental challenge of needing to understand where they are positioned as they move through the world. Towards this end, roboticists have extensively investigated solutions to the problem of simultaneous localization and mapping (SLAM), whereby systems must infer both a map of their surroundings and their relative locations as they navigate through space (Cadena et al., 2016). Considering that these same challenges face any freely moving cybernetic system, natural selection has similarly exerted extensive teleonomical (i.e., illusory purposefulness) optimization in this direction (Dennett, 2017; Safron, 2019b), so generating mechanisms for enabling wayfinding and situating organisms within environments where they engage in multiple kinds of adaptive foraging. Perhaps the most sophisticated of all biological SLAM mechanisms is the hippocampal-entorhinal system (H/E-S), whereby vertebrates become capable of both remembering where they have been, inferring where they are, and shaping where they are likely to go next.

Here, we argue that the development of the H/E-S represented a major transition in evolution, so enabling the emergence of teleology (i.e., actual goal-directedness) of various forms (Safron, 2021b), ranging from governance by expected action-outcome associations to explicitly represented and reflexively modellable causal sequences involving extended self-processes. We focus on the implications of SLAM capacities *via* the H/E-S, and of evidence that this functionality may have been repurposed for intelligent behavior and cognition in seemingly non-spatial domains. We propose that all cognition and goal-oriented behavior (broadly construed to include mental actions) is based on navigation through spatialized (re-)representations, ranging from modeling abstract task-structures to temporal sequences, and perhaps even sophisticated motor control *via* SLAM with respect to body maps. Indeed, we would go as far as to suggest that the ubiquity of implicit and explicit spatial metaphors in language strongly points to a perspective in which cognition is centered on the localization and mapping of phenomena within both concrete and abstract feature spaces (Lakoff and Johnson, 1999; Bergen, 2012; Tversky, 2019).

In these ways, we believe Generalized Simultaneous Localization and Mapping (G-SLAM) may provide enactive groundings for cognitive science within the principles of ecological rationality (Todd and Gigerenzer, 2012). That is, we adopt a perspective in which cognition is traced back to its ultimate origins, wherein rationality is understood in terms of adaptations for shaping animal behavior in ways that further evolutionary fitness. Such ecological and ethological connections further provide bridges to optimal foraging theory and (generalized) search processes as ways of understanding

cognition as a kind of covert behavior (Hills et al., 2013). While somewhat similar models of intelligence have been proposed (Hawkins, 2021), we suggest these other views may be somewhat misleading in neglecting to account for the central role of the H/E-S for realizing G-SLAM. In addition to providing an accurate viewpoint that grounds cognition in its cybernetic function as shaped over the course of evolution and development, G-SLAM will further allow rich cross-fertilization of insights between cognitive science and artificial intelligence. Given the particular functionalities enabled by the H/E-S, we propose this reverse-engineering project ought to be the central focus of cognitive science and machine learning, potentially constituting the most viable path forward towards realizing AI with advanced capacities for reasoning and planning (Bengio, 2017).

A thorough discussion of these issues is beyond the scope of a single manuscript. However, below we attempt to provide an overview of why we believe the G-SLAM perspective may provide a unification framework for cognitive science. First (in Section "LatentSLAM, a bio-inspired SLAM algorithm"), we review our work on biologically-inspired SLAM architectures for robotics. Then, we consider features of the H/E-S, including its functionality for localization and mapping in both physical and abstract domains. Finally, we discuss correspondences between features of SLAM and core aspects of cognitive functioning. We hope to explain how common principles may apply not only to the fundamental task of finding one's way to desired locations in physical space, but for thought as navigation through abstract spaces. While much of what follows will necessarily be under-detailed and speculative, in subsequent publications, we (and hopefully others) will explore these issues in greater detail as we attempt to explain fundamental principles in neuroscience and artificial intelligence, while simultaneously seeking synergistic understanding by establishing conceptual mappings between these domains (Hassabis et al., 2017).

In the following section, we provide a high-level overview of LatentSLAM, which is also treated in greater detail in (Çatal et al., 2021a,b). While we believe many of these technical details may be relevant for explaining fundamental aspects of high-level cognition, a more qualitative understanding of this content should be sufficient for considering the conceptual mappings we (begin to) explore in this manuscript (**Table 1**). Section "The Hippocampal/Entorhinal System (H/E-S)" then summarizes current views on the H/E-S and its functioning in relation to spatial modeling and cognition more generally. Finally, Section "G(eneralized-)SLAM as core cognitive process" draws parallels between understanding in machines (using LatentSLAM) and humans (considering the H/E-S) and propose G-SLAM as a unification framework for cognitive science and artificial intelligence.

We realize that this may be a challenging manuscript for many readers, with some portions focused on describing a robotics perspective, and other portions focused on cognitive/systems neuroscience. Indeed, this article emerged from an ongoing collaboration between roboticists and a cognitive/systems neuroscientist, which has been both rewarding and challenging in ways that demonstrate why this kind of interdisciplinary work is both desirable and difficult. One of our primary goals for this manuscript is to provide a rough-but-useful conceptual scaffolding (i.e., an initial partial map) for those who would attempt such cross-domain research. In this way, interested readers ought not be overly concerned if some of the content is found to be excessively technical relative to their particular background. However, we believe readers who follow through with exploring these suggested mappings (which we only begin to characterize) may be richly rewarded for those efforts.

In brief, G-SLAM can be summarized as follows:

1. It is increasingly recognized that the H/E-S may be the key to understanding high-level cognition.
2. Within the field of robotics, the H/E-S has been identified as having been shaped by evolution for the problem of simultaneous localization and mapping (SLAM) for foraging animals, and where these capacities appear to have been repurposed for navigating through other seemingly non-spatial domains.
3. We believe it would be fruitful to explicitly think of the core functionalities of SLAM systems and test whether these are not just reflected in the functioning of the H/E-S with respect to physical navigation, but with respect to other high-level cognitive processes as well.

If the H/E-S is the kind of gateway to high-level cognition that it is increasingly suggested to be (Evans and Burgess, 2020; George et al., 2021; McNamee et al., 2021), and if it can be well-modeled as having been selected for SLAM functionalities that were later repurposed, then we believe the difficulty of exploring the following material will more than repay the effort of attempting to make the journey. We also ask readers to note places where spatial language can be found, only some of which was intentional. Indeed, we take such linguistic spatializations as supporting evidence for the G-SLAM perspective, which perhaps may be overlooked by virtue of its very ubiquity (cf. fish not noticing water). This is not to say that all spatial cognition points to a SLAM perspective. Yet we believe such spatial mappings are notable in affording opportunities for localization and mapping with respect to such domains. We leave it up to the discernment of our readers to assess how far one can go with following such paths through conceptual spaces, which may not only provide new perspectives on familiar territories on minds, but may even make inroads into discovering how we may follow similar paths to the destination of creating artificial systems with capacities that were formerly considered to be uniquely human.

TABLE 1 Potential correspondences between LatentSLAM, cognitive psychological, and bio-computational phenomena.

LatentSLAM	Cognitive-psychological processes	Bio-computational processes
Mapping/graphing:	Inferring dimensions of feature spaces and relative locations of phenomena based on observations	Relations between hippocampal place cells for particular locations combined with entorhinal grid cells for multi-scale metric-affordance information
Localization:	Positioning specific phenomena (including the mapping and localizing system itself) within inferred feature spaces	Conjunction of hippocampal/entorhinal place/grid cells for positioning specific events within maps/graphs
Sensor and actuator uncertainty:	Perceptual (including mnemonic and imaginative) ambiguity	Body and world states are indirectly inferred based on partial information from noisy signaling systems
Views:	Visuospatial perception (as a function of actions)	Information from ventral and dorsal visual streams (and other modalities) organized according to egocentric perspectival reference frames (*via* posterior midline structures)
Proprioceptive poses:	Somatospatial perception (as a function of actions)	Frontal-parietal hierarchies over the somatomotor strip, with modeling/control potentially enhanced *via* explicit mapping of lateral parietal body schemas by other systems (e.g., midline structures coupling with the H/E-S)
Experience-map:	Structuring of episodic memory and imagination both informed by and informing visuospatial and somatospatial modalities	Transitions between hippocampal place fields entailing spatiotemporal trajectories for organisms (potentially including trajectories for important effector/sensor systems such as eyes and hands), both entrained by and entraining largescale cortical attracting states
Spatial landmark graphs:	Consciously-accessible representations of (salience-biased) spatial relations, potentially constituting our sense of space; semantic content of graph is based on actions and corresponding sensations as paths are traversed across/through these nodes	Hippocampal place fields as chained attractors, mutually entrained with cortex to orchestrate attracting states for population activity along reduced-dimensionality manifolds for both overt and covert action-perception cycles at and between these locations
Hierarchical generative model:	The processes by which a coherent stream of experience is generated and remembered with respect to both action and perception	A functional and algorithmic understanding of the brain as a hybrid machine learning architecture for predictive control of an embodied-environmentally-embedded agent
Fisher information metric:	The amount of information gained when traveling along a trajectory given a probabilistic generative model, wherein autonomous functioning is realized by minimizing discrepancies between predicted goal and present estimated states (*via* active inference); with respect to structure learning, the amount of "cognitive work" required to make sense of a domain	The amount of neural activity that must be expended to achieve adaptive cybernetic functioning in a given context, including with respect to constructing and refining world models entailed by patterns of effective connectivity
Accumulation of map uncertainty:	Deviations between models and that which is represented due to uncertainty with respect to cognition and latent world states	Deviations between likely patterns of neuronal attractor dynamics and their ability to orchestrate either overt or covert action-perception cycles (i.e., behaving or imagining) for autonomous functioning; cybernetic (and potentially thermodynamic) entropy for nervous systems
Loop-closures:	Events in which a familiar location in feature space is encountered with high confidence	High degrees of converging mutually consistent activity from the H/E-S and non-H/E systems
Graph-relaxation:	Assimilation of novel information into existing schemas *via* iterated distribution of updates across interconnected cognitive structures	Updating connectivity patterns to influence relative positioning of hippocampal place fields, potentially accompanied by largescale reductions in Hopfield energy
Node creation:	Accommodation of novel information *via* altering the structure of cognitive maps/graphs, potentially resulting in major updates to internal working (world) models with novel concepts	Creation of new place fields, involving various forms of (potentially neuromodulator-dependent) hippocampal plasticity, and/or establishment of new prefrontal attractors (i.e., patterns of canalized striatal-cortical loops)
Navigation:	Setting destinations in generalized space, which function as sources of prediction-error to be minimized through active inference; this may apply to the organism as a whole moving through (generalized) space, or to trajectories for parts of a system for which specific intentional control is warranted (e.g., directed ocular foveations or grasping/pointing movements), including with respect to spaces of a conceptual variety (e.g., spatialized time)	Predictive sweeps of activity across place fields from hippocampal maps (cf. successor representations), which can orchestrate largescale cortical attracting states (cf. equilibrium points) and thereby drive both system-internal self-organization (i.e., perceptual inference, imagination, and learning) and overt enaction, which in turn creates new sources of information to shape subsequent H/E-S dynamics

Please note, these cross-domain mappings are neither meant to be exhaustive nor definitive, but are instead intended to point in the direction of what a G-SLAM perspective might look like if more fully developed.

LatentSLAM, a bio-inspired SLAM algorithm

Simultaneous localization and mapping (SLAM) has been a long standing challenge in the robotics community (Cadena et al., 2016). For autonomous functioning, a robot must try to map its environment whilst trying to localize itself in the map it is simultaneously constructing (i.e., SLAM). This setup creates a kind of "chicken and egg" problem in that a well-developed map is required for precise localization,

but accurate location estimation is also required for knowing how to develop the map by which locality is estimated. This challenge is rendered even more difficult in that not only must the system deal with the seemingly ill-posed problem just described, but the inherent ambiguity of the environment is made even more difficult by sources of uncertainty from sensors and actuators. A fundamental challenge (and opportunity) with localizing and mapping is the detection of loop-closures: i.e., knowing when the robot re-encounters a location it has already visited. The challenge is due to the circular inference problems just described, and the opportunity is due to the particularly valuable occasion for updating afforded by the system having a reliable reference point in space. Such loop-closures have a further functional significance in allowing experiences to be bound together into a unified representational system where updates can be propagated in a mutually-constrained wholistic fashion, so providing a basis for the rapid and flexible construction and refinement of knowledge structures in the form of cognitive schemas that have both graph-like and map-like properties. With further experience, these schemata can then be transferred to the neocortex in the form of more stable adaptive action and thought tendencies, so forming a powerful hybrid architecture for instantiating robust causal world models (Hafner et al., 2020; Safron, 2021b).

SLAM has traditionally been tackled by Bayesian integration of sensor information within a metric map, typically expressed in terms of absolute distances and angles. In previous work, this amounted to keeping track of distances between the robot and various landmarks in the environment. Distance measurements were typically combined through Bayesian filtering, a principled way of combining heterogenous information sources through Bayesian inference. Modern successful metric SLAM solutions, however, combine lidar scans with the robots internal odometry estimate through Kalman filtering (Kalman and Bucy, 1961) into 2D or 3D occupancy grid maps (Mur-Artal et al., 2015; Hess et al., 2016). These occupancy maps (**Figures 1B,C**) keep track of locations of objects in the environment by rasterizing space and then marking certain grid locations as inaccessible—due to being occupied with physical obstructions—so creating a map that resembles what an architect would create to diagram a room (**Figure 1A**).

Variations on this scheme are popular and differ wildly, either substituting the integration algorithm or the type of metric map. A metric map is akin to a Cartesian grid with regular spacings. However, such spatial maps do not speak to the object identities within the space of interest, nor the particular relations between those objects. Thus, one of the downsides of using metric maps is that by extension all robotic reasoning must also happen on a metric level, any semantic information (i.e., the meaning of a certain cluster of grid-cell activations) needs to be added in later. Further, such metric spaces represent an instance of deviating from natural designs, as hippocampal/entorhinal system (H/E-S) mappings are not independent of the objects contained within these spaces, but instead induce distortions (e.g., expansions and compressions) of spatial relations, which are also modulated as a function of the salience of these entities for the organism/agent (Bellmund et al., 2019; Boccara et al., 2019; Butler et al., 2019).

Popular approaches for such spatiotemporal modeling use particle filters or extended Kalman filters as Bayesian integration methods (Thrun et al., 2005). Kalman filters are notable in that they allow for estimation based on a precision-weighted combination of probabilistic data sources, so allowing for synergistic power in inference and updating, which is also theoretically optimal in making use of all available data (weighted by relative certainty). As will be discussed in greater detail below, such integration may be implemented in the H/E-S *via* convergent activation in regions supporting high degrees of recurrent processing, such as the CA3 subfield of the hippocampus. However, not only does the H/E-S promote integrative estimation, but also pattern separation/differentiation *via* other subregions such as CA1, so allowing for attractors to take the form of sparsely-connected graphs—cf. hybrid continuous/discrete architectures based on Forney factor graphs and agent-designs based on independently controllable factors (Friston et al., 2017b; Thomas et al., 2017, 2018). Below we will also describe how such graph-like representations not only help to solve problems in navigating through physical spaces, but may also form a basis for the kinds of high-level cognition sought after in the domain of neurosymbolic AI (Bengio, 2017).

We do not internally represent the world in a metric map. For instance, none of our senses can naturally give us an accurate distance measurement. Neither are we very effective in following a metric description of a path. Hence, it makes more sense for minds like ours (and potentially for artificial agents) to represent a map intuitively as a graph-like structure (**Figure 1D**), where subsequent graph-nodes could represent subsequent high-level parts of the environment e.g., a node could represent a part of the environment containing a door at a certain rough location. Map traversal then becomes equivalent to the potentially more intuitive problem of graph-traversal or navigating between meaningful landmarks. Trajectories can then be expressed in terms of consecutive semantically meaningful directions. For example, the metrical path "move forwards 2 meters, turn 90 degrees clockwise and continue for 2 meters" could become "after going through the door go right towards the table." (Note: in vertebrate nervous systems, such forms of navigation could either be based on H/E-S graphs/maps, or occur *via* canalized striatocortical loops implicitly mapping states to actions, possibly with functional synergy, and also enhanced robustness (and thereby learnability) *via* degeneracy/redundancy.)

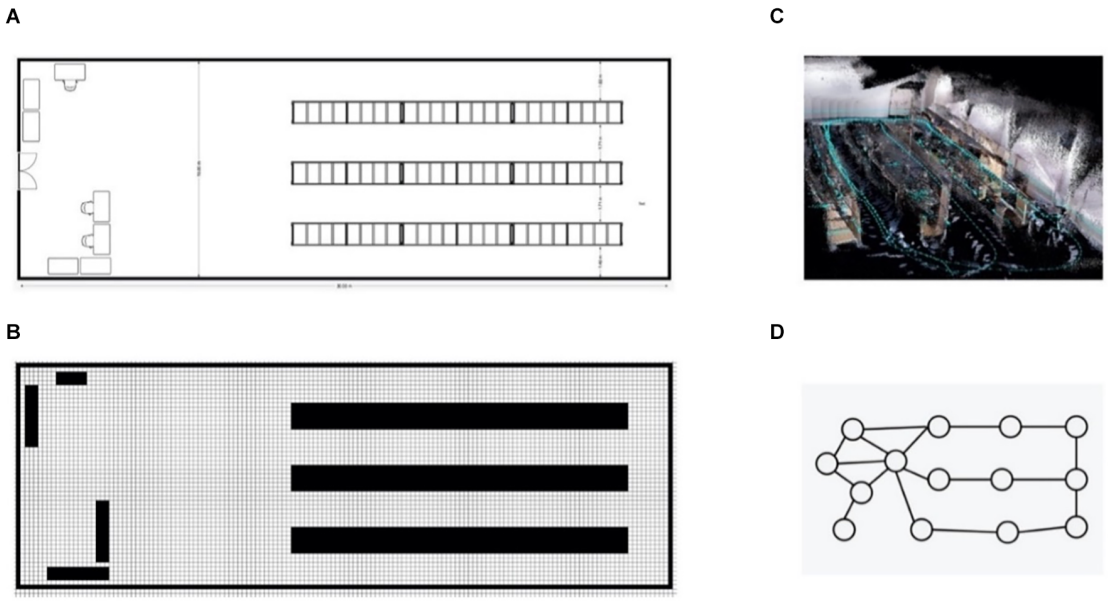


FIGURE 1
An overview of different map types, show-casing our robotics lab. Panel **(A)** gives an exact metric view of the room as drawn by an architect. Panel **(B)** shows the same map as a 2D grid map, to create this map from panel **(A)** the map was rasterized and untraversable terrain was filled into the granularity of a single raster cell. Pabel **(C)** shows the same room as an x, y, z mapping of red/green/blue values extracted from a RGBD camera. This 3D grid map was generated by moving the camera through the physical lab. Finally, panel **(D)** shows the lab as a sparse graph.

In LatentSLAM (Çatal et al., 2021a), we proposed a bio-inspired SLAM algorithm which tries to mimic this kind of intuitive mapping. With this architecture, we built topological, graph-based maps on top of a predictive model of the world, so allowing for separation of the low-level metric actions of the robot and high-level salient paths. Instead of using raw sensory data—or fixed features thereof (Milford et al., 2004)—directly as node representations, LatentSLAM learns compact state representations conditioned on the robot's actions, which are then used as nodes. This latent representation gives rise to a probabilistic belief space that allows for Bayesian reasoning over environmental states. Graph nodes are formed from trajectories on manifolds formed by belief distributions. That is, rather than utilizing static maps, our agents navigate through space by moving between landmarks based on expectations of which state transitions are likely to be associated with those kinds of percepts. As an underlying foundation, LatentSLAM adopts the Free Energy Principle and Active Inference (FEP-AI) framework to unify perception (i.e., localization), learning (i.e., map building) and action (i.e., navigation) as a consequence of the agent optimizing one sole objective: minimizing its (expected) free energy (Friston, 2010; Friston et al., 2017a). As will be described in greater detail below, we believe this is an apt description of thinking as the unfolding of a stream of consciousness, with a variety of somatic states being generated in various combinations as the agent perceives and imagines itself moving through space and time.

Representing the world in a graph

Graphs form a natural way of representing relations between various sources of information in a sparse and easily traversable manner. In LatentSLAM, such a structure is used to build a high-level map from agent experiences. This experience map contains nodes consisting of a *pose*, i.e., the agent's proprioceptive information, and a *view* distilled from the sensory inputs. Together, the pose and view of an agent specify its unique experience: a different view in the same pose gives rise to a new experience; likewise, the same view from a different pose also constitutes a novel experience. Views generally lie on some learned compact manifold as a compressed version of one or more sensory inputs, integrated and updated through time. Links between experiences in the graph indicate possible transitions between one experience and another.

Figure 2 provides a visual overview of how poses and views combine into an integrated experience map. The pose information allows the agent to embed the graph relative to the geometrical layout of the environment. In this case, the embedding is done in 2D-Cartesian space as the example shows a ground based, velocity-controlled mobile robot. Embedding the graph in a reference frame correlated with environment characteristics organizes observations in ways that greatly enhance inferential power, since this avoids combinatorial explosions with respect to under-constrained hypothesis spaces. That is, a given sensory impression could correspond to an

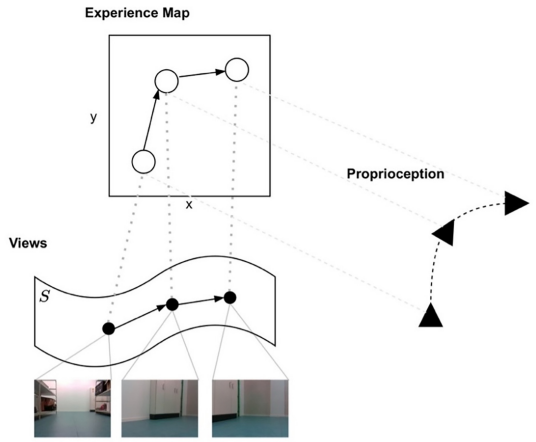


FIGURE 2
The formation of an experience-map out of views and proprioceptive poses. Sensory observations first need to be integrated into views to be compared to existing experiences from the graph. The shown graph is embedded in a Cartesian reference frame extracted from the proprioceptive information.

unbounded number of world states (e.g., something may be big and far away, or small and nearby), but coherent perspectival reference frames allow for likely causes to be inferred by mutually-constraining relevant contextual factors.

Experience map

The experience map (or graph) provides a high-level overview of the environment. Each node in the map represents a location in the physical world where the robot encountered some interesting or novel experience. These positions are encoded in poses in a spatial reference frame, e.g., a 2D-Cartesian space, whilst the experiences themselves are expressed as implicit representations of corresponding sensory observations. When view representations change according to distances to known landmarks, this setup resembles the approach described in the classical graphSLAM algorithm (Thrun and Montemerlo, 2006). Note that the seminal work on graph-based experience maps (Milford et al., 2004) also used an embedding of sensory observations into a lower dimensional space. However, in contrast to our approach, these mappings were deterministic and fixed for all observations.

The graph is embedded in, as opposed to being expressed in, a spatiotemporal reference frame, meaning that over time stored (or inferred) poses on the map are likely to exhibit deviations from their initial recorded values as they are progressively updated. Loop-closure events trigger a graph-relaxation phase wherein current graph nodes are re-positioned to take into account the unique opportunity accompanying the closing of the loop (i.e., the creation of a closed system of node linkages allowing for updating of the entire graph through energy minimization, accompanied by more confident location-estimation through experience-trajectory converging on known landmarks). This relaxation not only affords opportunities for map refinement, but it is also necessary due to the accumulation in pose errors from odometry drift. Wheel slippage, actuator encoder errors, and other similar effects amount to a continual increase in the uncertainty of the pose estimate. These sources of error/noise are part of what makes loop-closure such a hard problem in general. However, the loose embedding of pose information in the graph (combined with associated views) allows the map building to become robust to sensor and actuator drift, thereby maintaining a consistent map of the environment.

Views

LatentSLAM probabilistically learns views from sensory observations by incorporating the action trajectories from which they are generated, which differentiates our architecture from similar algorithms (Milford et al., 2004). The agent keeps track of a sample of the current belief distribution over states, which gets updated at each time-step into a new belief through variational inference. This sample constitutes either the current agent view, or a sensory-decoupled (or imagined) estimate of the environment from the latent space of the agent's generative model. At each time-step, the agent inputs a conjunction of the current action, sample, and current observation into its generative model. This world model then generates a new state belief distribution based on the current state sample, which functions as a source of predictions for a predictive coding perceptual architecture. At training time, the generative model is tasked with predicting future observations based on previous recordings of trajectories through the environment.

Proprioception

An agent needs a principled way of keeping track of its estimated *pose* in the local environment. That is, an agent needs a coherent way to integrate changes in its local pose according to some local reference frame. In this form of proprioception, agents can estimate the effects of certain actions on local pose information relative to adjacent portions of its environment. This aspect of embodiment is essential in enabling consistent mapping and localization through challenging terrains.

In LatentSLAM this is handled through the low-level generative model on the one hand, and the pose continuous attractor network (CAN) on the other hand. The generative model allows for reasoning in terms of how actions affect views: i.e., it reduces the pose to an implicit part of the latent state representation. The CAN, however, leaves pose estimation as an explicit part of the greater LatentSLAM model. It integrates successive pose estimates through time in a multidimensional grid representing the agent in terms of internally measurable

quantities. In the case of a ground-based mobile robot these quantities would be the expected difference in *x,y* pose and relative rotation of the robot over the z-axis. Hence, for a ground-based robot the CAN would be expressed as a 3D grid, that wraps around its edges. Sufficiently large displacements along the x-axis of this grid would teleport the pose estimate back to the negative bound of the same axis. This to accommodate for traversing spaces that are larger than the number of grid cells in the CAN. The pose estimate in the CAN is represented as an activation per grid cell, the value of which determines the amount of belief the model gives to the robot being in this exact relative pose. Multiple grid cell locations can be active at any given time, indicating varying beliefs over multiple hypotheses. The highest activated cell indicates the current most likely pose. Cell activity is generated in two ways: activity is added (or subtracted) to a cell through motion and the current proprioceptive translation thereof in terms of grid-cell entries; alternatively, activity may be modified through view-cell linkage. When a view is sufficiently different from others it gets added to the experience map together with the current most likely pose. This mechanism in turn allows experiences, when encountered, to add activation into the CAN at the stored pose estimate. This process can shift, and often correct, the internal pose estimate of the agent, allowing it to compensate for proprioceptive drift.

This conjunction of views and poses has notable parallels with neural representations decoded from respective lateral and medial entorhinal cortices (Wang C. et al., 2018), which constitute the predominant source of information for the hippocampal system (i.e., the experience map). It is also striking that the self-wrapping representational format for LatentSLAM poses/views recapitulates the repeated metric-spacing observed for entorhinal grid cells, whose location invariance may potentially provide a basis for knowledge-generalization and transitive inference across learning epochs and domains (Whittington et al., 2022). We believe that such correspondences between naturally and artificially "designed" systems constitutes strong evidence in support of a SLAM perspective for understanding the H/E-S.

A hierarchical generative model

The entirety of the LatentSLAM framework can be understood mathematically in terms of a hierarchical generative model (**Figure 3**; Çatal et al., 2021b).

There are two distinct levels of reasoning, each using their own generative model to explain the dynamics of the environment at the corresponding level of abstraction. As the generative models are stacked, the higher-level model takes the states from the lower level as observations, while the lower level observes the actual environment through the agents' sensors. Each separate generative model can be seen mathematically as representing the joint probability $p(\tilde{o}, \tilde{s}, \tilde{a}) = p(a_0)p(s_0)p(o_0|s_0)\prod_{t=1}^{T} p(s_t|s_{t-1}, a_{t-1})p(o_t|s_t)$, with o relevant observations at each level; s state description, views or locations; and a possible actions at each level (either displacements in the environment or node transitions). These models only consider the generative process up until some future time horizon T. The exact instantiation of the joint probability and corresponding posterior distributions differ between each level of the hierarchy; interested readers are referred to Çatal et al. (2021b) for a more thorough description of this kind of model, and some extra details are provided in the "**Appendix**".

Action and state inference, that is finding suitable instantiations of the posteriors $p(a_t|s_t)$ and $p(s_t|s_{t-1}, o_t, a_{t-1})$ is achieved through Active Inference as understood in the context of the Free Energy Principle (FEP-AI; Friston et al., 2017a). In FEP-AI, intelligent agents are governed by predictive models that attempt to minimize variational free energy through updating of internal beliefs and modification of external states through enaction (hence, active inference). When implementing similar mechanisms in artificial agents such as robots, inference is amortized—cf. planning as inference *via* memorization of successful policies (Gershman and Goodman, 2014; Dasgupta et al., 2018)—through training variational auto-encoders (VAEs) with objective functionals that minimize (variational) free energy. The model consists of three neural networks, with each representing a conditioned probability distribution that outputs different multivariate Gaussian distributions based on differing inputs. These inputs can take the form of different sensor modalities such as lidar or camera; or they might be actions depending on the flow of information between neural networks.

State inference emerges naturally from the neural network architecture and training method. Active inference, however, leverages the trained network to create a set of imaginary trajectories from which optimal action sequences can be selected through expected free energy minimization. The model is trained on a free-energy objective functional, wherein it is tasked with minimizing Bayesian surprise—in the form of KL divergence—between prior and posterior estimates on the state. In this hierarchical generative model, there are two sources of information flowing in two directions at any given time. Sensory observations flow upwards from the real world through the lower-level pose-view model towards the higher-level mapping model. Predictions flow in the opposite direction, originating in the higher-level mapping model and flowing down into the environment through the predicted actions in the lower-level pose-view model.

Bottom-up sensory streams

The agent observes the world through sensors as it moves around the environment. At the lower-level of the generative model, the agent actively tries to predict future incoming sensory

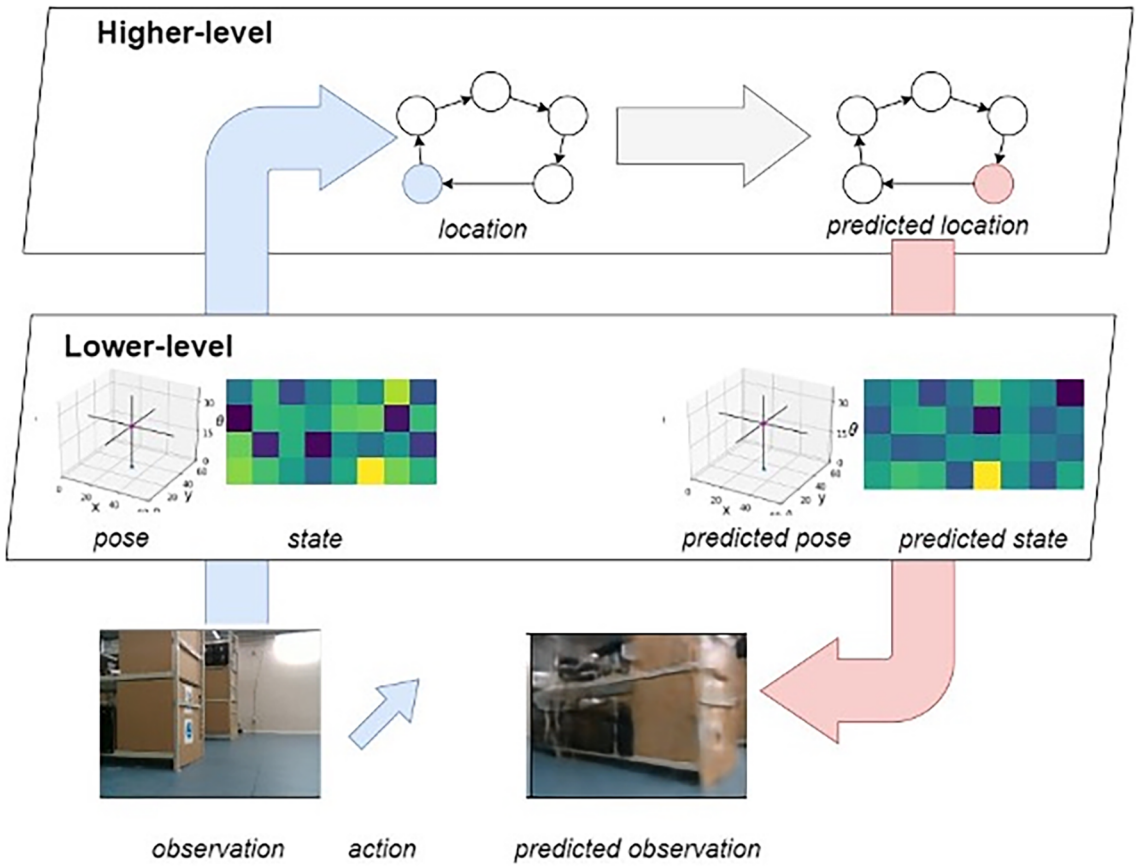


FIGURE 3
Overview of the hierarchical generative model. Highlighted in blue is the bottom-up sensory stream, and in pink the top-down prediction stream. As the agent moves about, it alternates between these two modes. On the one hand it will infer state information from the observations, and on the other hand it will predict future observations from inferred states.

observations (Figure 3, blue arrow indicating informational flow). The agent actively abstracts away distractor elements in the observations as every observation gets encoded into a latent vector (i.e., views). As this encoding is generated from actions, observations and the previous latent state, the model considers the effects that history and actuation (or enaction) have on the environment. The abstracted view then gets fed into the higher-level mapping model which actively predicts the next experience from the previous one, taking into account the way the agent is presently traversing the experience graph and its current view.

Top-down prediction streams

At the same time, decisions flow down from the higher-level to the lower-level of the generative model (Figure 3, red arrow indicating informational flow). As a new navigational goal is set, the desired trajectory through the experience map is generated. Each node transition denotes one or more displacements in the real environment. While traversing the graph, the agent sets the views associated with the visited nodes as planning targets for the lower-level model. At the hierarchically higher level, the agent samples multiple state estimates from the current belief distribution over states and leverages the predictive capabilities of the generative model to envision possible outcomes up until some fixed planning horizon (Friston et al., 2021). From all these imagined future outcomes, the optimal one is selected after which the process repeats itself until the target view and pose are met. In turn the next node in the map trajectory is used to generate a new lower-level planning target.

Creating the map

As mentioned earlier, once an agent encounters a sufficiently different experience, a new node is inserted in the experience map with the current view and pose. This process results in an ever-growing map of the environment as the agent explores the world. Hence, there needs to be a principled way to determine whether a view is new or is already known to the agent. As with

many such problems, the solution presents itself in the form of a distance function in some well-defined mathematical space. A well-chosen distance function will allow the agent to not only build a consistent map of its environment but also account for loop-closure events.

Distance functions

Many SLAM algorithms use the Euclidian distance between poses to determine whether the current observation and pose are known in the map or represent some novel experience. However, due to the inherent drift in proprioception in many real-world scenarios, often this distance metric between poses and/or observations is not enough. Alternatives present themselves depending on the form of the probabilistic framework upon which the algorithm is based.

As described in Section "A hierarchical generative model", LatentSLAM learns a latent state space manifold over sensory inputs (i.e., camera images). This enables the agent to not only evaluate Euclidian distances between poses, but also distances between two sensory inputs in the latent statistical manifold. To evaluate distances inside the manifold we need an appropriate distance measure. One notable candidate is the Fisher information metric (Costa et al., 2015), which represents informational differences between measurements. In our context, this means that two measurements are only encoded in different nodes of the experience map when there is sufficiently more information in one compared to the other. For example, moving in a long, white hallway with little texture will not yield a change in information in the latent manifold, hence this will be mapped on a single experience node. Only when a salient feature appears, for example a door, there will be enough sensory information to encode a new experience. In such a scenario however, methods building a metric map will likely fail as it is impossible to accurately track one's position in a long, textureless hallway.

Note how the Fisher information metric is also related to the free energy minimization objective used for manifold learning. Concretely, if we take $KL[x||x+\delta x]$ with x a probability distribution and $x+\delta x$ a distribution close to x we get that if $\delta x \to 0$ then $KL[x||x+\delta x] \to \frac{1}{2}F(x)(\delta x)^2$. In other words, for infinitesimally small differences between distributions the KL divergence approaches the Fisher information metric (Kullback, 1959). This can be interpreted as integrating the agent's Bayesian surprise over infinitesimal timesteps to measure the "information distance" traveled.

However, since the Fisher information metric and KL divergence do not have closed form solutions for many types of probability distributions, we use cosine similarity between the modes of the distribution as a numerical stable approximation function. Therefore, LatentSLAM evaluates *information* differences between experiences instead of differences in exact environmental observations.

Node creation and loop-closures

When a salient landmark is identified, but the agent cannot find a single node in the graph which matches closely enough with the current view or pose, a new node must be inserted in the graph. Alternatively, if the current experience matches both on pose and view, a loop-closure is registered, but the agent leaves the map as is. In order to determine whether two experiences match, LatentSLAM uses a matching threshold θ. Both the pose and view of an experience is matched to experiences stored in the map. **Figure 4** gives a visual overview of the various possible matching cases. If neither view nor pose match with any possible stored view or pose, a new experience is created and inserted into the map, as is shown in panel A. When the view and the pose both match, a loop-closure has occurred and the current experiences shifts to the stored experience, at which point a graph-relaxation phase is initiated. If the current observed experience matches with a stored experience further along the path, a relocation is required, and the estimate is shifted further along the path in the graph. Finally, if the current pose estimate matches a stored experiences pose, but does not find the corresponding matching view, a new node is inserted at the same location. This allows the agent to keep track of varying views of the same landmark throughout the map.

Graph-relaxation

As nodes are inserted throughout the graph, each new pose observation is subjected to sensor drift, leading to increasing errors for remembered poses. To address this issue, whenever a loop-closure event is encountered, graph-relaxation is applied to the experience graph. The algorithm treats every node in the graph as being connected with its neighbors as if suspended by weighted springs. The strength of each spring is related to the pose distances between the nodes. Then the algorithm reduces the total "energy content" of the graph by shifting the poses in such a way that the sum of the forces is minimized. This approach is similar to graph-relaxation in similar SLAM algorithms (Thrun et al., 2005; Thrun and Montemerlo, 2006). Graph-relaxation has the effect of morphing the shape of the pose embedding of the map to reflect the actual topology of the environment.

Setting the threshold

Because the matching threshold has a significant impact on the shape and content of the map, it is one of the more important hyper parameters of LatentSLAM. For every environment there is an optimally tuned threshold parameter $\theta^{\star}$. A matching

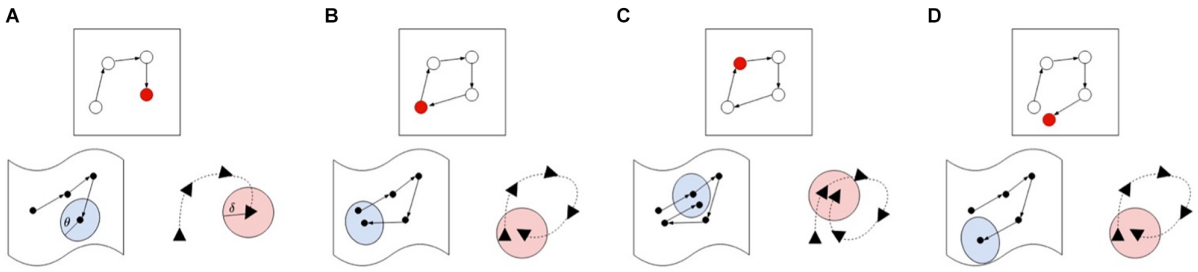


FIGURE 4
Different cases for illustrating the map updating procedure. For each case we show the map (top), pose (bottom right), and views (bottom left) in their own respective spaces. The current active map node is always indicated in red and the current pose or view value is the final one in the sequence. In case **(A)**, the agent encounters a new experience which is not within the threshold boundary of both the poses and views, so a new node is inserted into the map. Case **(B)** demonstrates a loop-closure event, where both the pose and view are within their respective thresholds, blue indicating the area pose information demarcated by its threshold θ, pink indicating the area covered by the view threshold. If both view and pose are within the threshold boundary (blue and pink) of the next node (case **C**), the estimate is shifted to the next node, skipping the current node in the graph. Finally, case **(D)** shows a matching pose without a matching view, requiring a new node insertion in the map.

threshold much lower than this optimal value will result in a mapping procedure with almost no loop-closure events. The map will contain every tiny permutation in views and poses as a separate node and will be insufficient in countering odometry drift. Conversely, if the threshold is set much higher than $\theta^{\star}$, the mapping procedure will lump everything together in a small cluster of nodes. **Figure 5** provides a visual example of the effects of the matching threshold on the resulting map.

Navigation

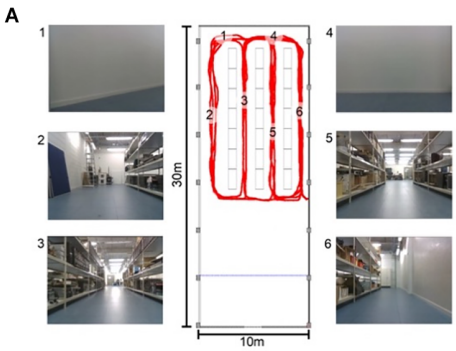
Navigation is achieved through a dual process of first selecting nodes in the higher-level experience map, and then setting the node-views as targets for the active inference based lower-level action planner. In the first phase a path is generated through the graph connecting the current node and the target node. The final node is selected based on the visual reconstruction of the stored view. That is, the user of the system selects the view they want the system to have at a certain place. Once an experience trajectory is found, the agent can start acting in the environment. As each consecutive experience node is separated from its neighbors by a finite set of actions, a sequence of target views are extracted from the trajectory, forming the imaginary trajectory the agent may (approximately) bring about through overt enaction. The (active inference based) lower-level generative model is then capable of filling in further gaps between imagination and reality through additional planning.

At each step, the agent takes into account its current view and imagined trajectory up until the next target view. This imagination process leverages the learned intricacies and dynamics of the environment to compensate for the potential stochasticity in the interaction. Once a suitable trajectory is imagined at the higher-level, the agent enacts the first step of the trajectory, after which the lower-level planning process is repeated. These step-by-step transitions through the environment make the agent more robust against unexpected changes in the environment, which it might not have captured during model training.

Crucially, imagined trajectories are scored using a common objective functional of expected free energy, both on the higher level of proposed paths through the experience map/graph, as well as on the lower level of inferring actions capable of transitioning the agent between nodes (Çatal et al., 2021b). That is, trajectories are more likely to be selected if they bring the agent towards preferred outcomes and/or resolve uncertainty about the environment. Hence, action selection comprises a trade-off between instrumental value and epistemic value, which are naturally balanced according to a singular criterion of variational free energy. To provide an example in navigation, this tradeoff between the extrinsic value of realizing prior preferences and the intrinsic value of novel information could respectively manifest as either selecting a safer route *via* well-recognized landmarks or instead taking an unknown (but potentially shorter) path through a dark forest. Further, the discovery of such shortcut paths through space speaks to the kinds of flexible inference and learning that first motivated construals of the hippocampal system in terms of cognitive maps (Tolman, 1948), and in a G-SLAM context could be thought of as a way of understanding a core aspect of intelligence in the form of creative insight. And in the context of AI, such creative cognition may afford the creation of much sought after capacities for powerful inferences and one-shot learning in novel situations, which if realized could greatly enhance autonomous functioning.

Limitations and future directions

There are several limitations with the current implementation of LatentSLAM. First, the experience



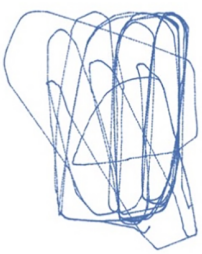


FIGURE 5
(A) Metric map of our lab environment, with some example camera views at the marked locations. The views at different locations (i.e., 3 and 5 or 1 and 4) appear very similar, making this a hard environment for visual SLAM. Panels **(B–D)** show three possible mappings of the trajectory shown in red in panel **(A)**. **(B)** With a well-tuned threshold θ^*, our LatentSLAM algorithm recovers a topological map of the environment, clearly separating the four different aisles. **(C)** If the threshold is too stringent ($\theta \ll \theta^*$), loop-closure events are not detected, as every view is seen as unique, and the map becomes incorrect as proprioception errors (the main source of mapping errors) add up. **(D)** When the threshold is too relaxed ($\theta \gg \theta^*$), similar looking aisles are mapped onto each other due to false positive loop-closures.

graph is incapable of merging nodes with similar views and approximately similar poses into a single unified stochastic node. This in turn leads the algorithm to generate an increasing number of nodes for each pass through a single location. Second, the lower-level planning is limited to the sequence length encountered during training, and as such the model is incapable of imagining coherent outcomes beyond this time horizon. This brings us to a potentially substantial limitation of LatentSLAM, in that the lower-level generative model needs to be pre-trained on the types of observations it can encounter in the environment. That is, when the target views are unknown, imaginative planning may be required wherein agents visualize an assortment of potentially rewarding (counterfactual) action-outcome pairings. Going forward, we aim to alleviate these constraints by adapting the training procedure to accommodate online learning, allowing the agent to learn to imagine whilst exploring (Safron and Sheikhbahaee, 2021), which may be understood as a kind of deep tree search through policy space *via* Markov chain Monte Carlo sampling (Dohmatob et al., 2020; Friston et al., 2021), with potentially relevant insights obtainable from advances in Bayesian meta-reinforcement learning (Schmidhuber, 2020).

To extend the biological fidelity (and potential functional capacities) of our architecture, we intend on attempting to recapitulate particular empirical phenomena such as the specific conditions under which new place fields are introduced or pruned away in mammalian nervous systems. For example, the insertion of environmental barriers or encountering corridors leading to identical rooms may induce duplication of sensory views at different locations, which may speak to the phenomenon of place-field duplication—which in a LatentSLAM context would involve node creation (Lever et al., 2009; Spiers et al., 2015)—yet where these representations may also disappear with further learning. This kind of pruning of nodes—potentially involving "artificial sleep"—could be a valuable addition to latent SLAM's functionality, and may potentially be understood as an instance of Bayesian model reduction with respect to structure learning (Friston et al., 2019), so providing another means by which capacities for creative insight (in terms of discovering more elegant models) may be realized in AI.

With respect to these particular phenomena involving challenging ambiguous situations, we may speculate that highly-similar-but-subtly-different pose/experience map combinations could represent instances associated with high levels of prediction-error generation due to a combination of highly precise priors and contradictory information. Speculatively, this could be understood as an example of "hard negative mining" from a contrastive learning perspective (Mazzaglia et al., 2022). As will be described in greater detail below, such highly surprising events may be similar to experiences of doorway or threshold crossing, and may trigger the establishment of event-boundaries *via* frame-resetting and spatial-retiling. Speculatively, the assignment of particular content to particular rooms in "memory palaces" could be understood as a necessary part of the art of remembering due to this phenomenon potentially interfering with semantic "chunking" (or coherent co-grounding). In attempting to apply LatentSLAM to cognition more generally, it could potentially be fruitful to look for generalizations of these phenomena with respect to seemingly non-spatial domains, such as with respect to creativity and insight learning problems in human and non-human animals.

Finally, and with further relevance to realizing capacities for imaginative planning and creative cognition, we will attempt to include phenomena such as sharp-wave ripples and forward/reverse replay across hippocampal place fields (Ambrose et al., 2016; de la Prida, 2020; Higgins et al., 2020; Igata et al., 2020), which have been suggested to form a

means of efficient structural inference over cognitive graphs (Evans and Burgess, 2020). With respect to our goal-seeking agents, forward replay may potentially help to infer (and prioritize) imagined (goal-oriented) trajectories, and reverse replay may potentially help with: (a) back-chaining from goals; (b) increasing the robustness of entailed policies *via* regularization, and (speculatively), and (c) allowing for a punishment mechanism *via* inverted orderings with respect to spike-timing-dependent-plasticity. In these ways, not only may a G-SLAM approach allow for deeper understanding of aspects of biological functioning, but attempting to reverse engineer such properties in artificial systems may provide potentially major advances in the development of abiotic autonomous machines.

The hippocampal/entorhinal system (H/E-S)

The hippocampal/entorhinal system (H/E-S) represents a major transition in evolution (Gray and McNaughton, 2003; Striedter, 2004), with homologs between avian and mammalian species suggesting its functionality becoming established at least 300 million years ago (Suryanarayana et al., 2020), with some of its origins potentially traceable to over 500 million years in the past with the Cambrian explosion (Feinberg and Mallatt, 2013), and potentially even earlier. It may be no overstatement to suggest that the H/E-S represents the core of autonomy and cognition in the vertebrate nervous system, with similar organizational principles enabling the potentially surprising degrees of intelligence exhibited by insects (Ai et al., 2019; Honkanen et al., 2019).

While their precise functional roles continue to be debated, the discovery of hippocampal place cells and entorhinal grid cells was a major advance in our understanding of how space is represented in the brain (O'Keefe and Nadel, 1978; Hafting et al., 2005). Similarly important was the discovery of head direction cells in rats, which were found to activate according to moment-to-moment changes in head direction (Sharp PE, 2001). Place cells have been modeled as representing a "predictive map" based on "successor representations" of likely state transitions for the organism (Stachenfeld et al., 2017), and grid cells have been understood as linking these graphs (or Markov chains) to particular events happening within a flexible (multi-level) metric tiling of space, so allowing for estimates of locations *via* path integration over trajectories. While we need not resolve the precise correspondences between these cell types here, there are intriguing developmental observations of place cells acquiring more mature functioning prior to grid cells, both of which likely depend on head-direction cells for their emergence (Canto et al., 2019; Mulders et al., 2021). In other contexts, place-specific cells have been found to index temporal sequence information, potentially functioning as "time cells" (Pastalkova et al., 2008), so providing a further means by which the H/E-S may provide foundations for coherent sense-making and adaptive behavior through the spatiotemporal organization of organismic information (Eichenbaum, 2014; Umbach et al., 2020).

In addition to place, time and grid cells, a variety of additional specialized cell types have been observed in the H/E-S. While it was previously assumed that these features represent innate inductive biases (Zador, 2019), increasing evidence suggests these specialized cell types may arise from experience-dependent plasticity, including models with similar architectural principles to the ones described here. In recent work from DeepMind (Uria et al., 2020), a recurrent system was used to predict sequences of visual inputs from (the latent space) of variational autoencoders. A natural mapping from egocentric information to an allocentric spatial reference frame was observed, including the induction of specialized units with response properties similar to head direction, place, band, landmark, boundary vector, and egocentric boundary cells. Similar results have been obtained with the Tolman-Eichenbaum machine (Whittington et al., 2020), including demonstrations of reliable cell remapping, so enabling transfer learning across episodes with the potential for the creative (re-)combination of ideas and inferential synergy. Other intriguing work on the emergence of specialized H/E-S functions through experience comes from work on "clone-structured cognitive graphs", where various aspects of spatial maps are parsimoniously formed as efficient (and explanatory) representations of likely state transitions through the duplication and pruning of nodes in a dynamically-evolving sequence memory (George et al., 2021). While this evidence suggests a potentially substantial amount of experience-dependence in the emergence of the "zoo" of specialized neurons for spatiotemporal navigation, the development of these features still involve clear innate inductive biases (Zador, 2019). Specifically, specialized pathways ensure that the H/E-S receives neck-stretch-receptor information from the mamillary bodies and yaw/pitch/roll information from the vestibular apparatus (Papez, 1937; Wijesinghe et al., 2015), so providing bases for sensor-orientation with respect to head-direction and thereby the foundations of egocentric perspective.

H/E-S as orchestrator of high-level cognition

While the association of the hippocampus with autobiographical and declarative memory is well-documented (MacKay, 2019), the H/E-S is increasingly being recognized as foundational for cybernetic functioning on multiple scales. A more thorough understanding of the principles governing the H/E-S and its interactions with the rest of the brain may allow us to understand how such sophisticated cognition and behavior is demonstrated by biological organisms

(Todd and Gigerenzer, 2012). Even more, such knowledge may also allow us to find ways of reproducing these functionalities in artificial intelligences.

The hippocampus is usually described in terms of a "trisynaptic circuit" (Andersen, 1975), with multiple specialized subsystems that interact with functional synergy. The dentate gyrus is the primary input area to the hippocampus from entorhinal cortex, with densely packed cells for pattern separation, so allowing for multiple separable/orthogonal representations. Much of this information then feeds into CA3, characterized by highly recurrent circuits with tight loops for dynamic pattern completion. This information is then routed to CA1, characterized by sparse and stable representations, representing the primary output area of the hippocampus and interface with the rest of the brain. Taken together, the subfields of the hippocampal complex allow multiple sources of information to be not just independently stored in memory, but also creatively combined within and across experiences, so affording powerfully synergistic functionalities such as transfer learning and generalizable knowledge. Intriguingly, some evidence suggests that humans might be unique in exhibiting less pattern separation in their hippocampal subfields, potentially contributing to—and possibly being a function of—cognition involving high degrees of abstraction/invariance (Liashenko et al., 2020; Mok and Love, 2020; Quiroga, 2020).

Indeed, the functional properties enabled by the H/E-S represent the state of the art in machine learning for real world applications such as autonomous vehicles and artificial intelligences attempting to realize higher-order reasoning abilities (Ball et al., 2013; Bengio, 2017; Hassabis et al., 2017; Kaplan and Friston, 2018; Shang et al., 2019; Eppe et al., 2020; Greff et al., 2020; Parascandolo et al., 2020; Shamash et al., 2020; Friston et al., 2021). This is a bold claim for a system that might be describable as an association machine or spatial mapper, which when lesioned tends to leave much of higher-order intelligence intact. However, closer inspection of hippocampal patients reveals its essential contributions to complex reasoning, emotion, and general behavioral flexibility (MacKay, 2019). It should also be kept in mind that while someone might be able to maintain certain functions after losing a system in adulthood—as this functionality may become distributed throughout the rest of the brain with experience—the congenital absence of a working H/E-S might be a wholly different manner, potentially precluding the bootstrapping of any kind of sophisticated cognition or coherent world modeling whatsoever (Safron, 2021a). Further, principles of association may be surprisingly powerful if they are capable of representing specific relational structures as particular graphs/networks, which are increasingly being recognized as powerful learning and inferential systems (Gentner, 2010; Zhou et al., 2019; Crouse et al., 2020). Some have even suggested that the mapping abilities of the H/E-S may provide bases for a potential core functionalities associated with conscious processing in the form of "unlimited associative learning" (Birch et al., 2020), in which knowledge may be flexibly aggregated across experiences (Mack et al., 2018, 2020; Mok and Love, 2019)—cf. transfer and meta-learning (Wang J. X. et al., 2018; Dasgupta et al., 2019; Kirsch and Schmidhuber, 2020). The central role of the H/E-S for higher-order cognition is further understandable in light of the fact that many (and possibly most) aspects of intelligence can be described as search processes (Conant and Ashby, 1970; Hills et al., 2010), which might be even more clearly apparent if we think of the possibility of spatializing abstract domains such as complex feature spaces (Eichenbaum, 2015; Whittington et al., 2018), or even time (Howard, 2018; Gauthier et al., 2019).

The H/E-S represents both the developmental foundation and functional apex of the cortical hierarchy (Hawkins and Blakeslee, 2004; Barron et al., 2020). In predictive processing models of the brain—e.g., the variational autoencoder framework described here—observations not predicted at lower levels eventually reach the entorhinal cortex and hippocampus. We propose the H/E-S allows these high-level prediction-errors to be temporarily encoded and organized with spatiotemporal and abstract relational structure for informational synergy. Indeed, on a high-level of abstraction, the H/E-S can be considered to be a kind of Kalman variational autoencoder that combines heterogeneous forms of (precision-weighted) information for SLAM in generalized state/phase space (Fraccaro et al., 2017; Zhang et al., 2017). Alternatively framed, the cortical predictive hierarchy can be viewed as hierarchical Kalman filtering all the way up and all the way down (Friston, 2010). Along these lines, it is notable that the H/E-S itself may operate in a manner that reflects more general principles of cortical predictive processing. With canonical microcircuits for predictive coding, predictions are associated with deep pyramidal neurons and alpha/beta frequencies, and prediction-errors are associated with superficial pyramidal neurons and gamma frequencies (Bastos et al., 2012, 2020). Consistently with the H/E-S involving predictive processing, novel information (i.e., prediction errors) induce activation of superficial pyramidal neurons for entorhinal cortex, dentate, and CA3, and recollection (i.e., predictions) are associated with activations in deep pyramidal neurons for CA1 and entorhinal cortices (Maass et al., 2014). Also consistently with a predictive processing account, another study observed superficial place cells in CA1 responding (*via* a rate code) in cue poor-environments, and deep pyramidal neurons responding (*via* a phase code) in cue-rich environments, where we might respectively expect either prediction-errors or predictions to predominate (Sharif et al., 2020).

From a predictive coding perspective, the hippocampus is a strange kind of cortex, not only because of its particular cytoarchitectonic properties (e.g., 3 vs. 6 layers), but also because of its connectomic centrality. Some proposals have suggested that memory recall may arise from "fictive prediction errors" (Barron et al., 2020)—a perhaps somewhat counter-intuitive

suggestion, in that the hippocampus is considered to be the top of the cortical heterarchy, and hence would be expected to only provide descending predictions—so providing a source of training signals for optimizing generative models of the world without sensory data, as well as affording stimulus-independent learning and imaginative planning. This is consistent with work from DeepMind in which the hippocampus is described as operating according to principles of "big loop recurrence", where its outputs can be recirculated as inputs for offline learning and counterfactual processing (Koster et al., 2018). Indeed, the H/E-S may not only provide sources of predictions for the neocortex, but potentially prediction-errors for itself, possibly by parameterizing simulations from cortical generative models (Higgins et al., 2020). Further, recent evidence regarding episodic memory formation and retrieval suggests that interactions between cortex and the H/E-S may reflect the roles of various frequency bands in predictive coding, or "routing" (Griffiths B. J. et al., 2019; Bastos et al., 2020). In this work, neocortical alpha/beta (8–20 Hz) power decreases reliably correlated with subsequent hippocampal fast gamma (60–80 Hz), and hippocampal slow gamma (40–50 Hz) power, potentially indicative of a trading off between predictions and prediction errors. However, this is somewhat different than the standard predictive coding account attributed to the cortex more generally, in that gamma frequency involvement may support the aforementioned idea that hippocampal reactivation of memories involve "fictive prediction errors" (Barron et al., 2020), rather than a suppressive explaining away.

In contrast to other slow rhythms, hippocampal theta oscillations may indicate enhancement of observations *via* cross-frequency phase coupling (Canolty and Knight, 2010), potentially providing a basis for high-level action and attentional selection. Along these lines, the ability of theta-oscillations to select and orchestrate cortical ensembles at gamma frequencies may provide a role for the hippocampal system as a comparator, enabling contrasting between percepts, whether based on observations or imagination (Safron, 2021b). Opposite phase relations between CA1 and CA3 (Tingley and Buzsáki, 2018) are suggestive, potentially indicating both a kind of predictive coding within the hippocampal system, and possibly also instantiating and orchestrating the formation and contrasting of corresponding cortical ensembles as alternating phases of duty cycles for theta oscillations (Heusser et al., 2016; Kunz et al., 2019). Indeed, the entertainment of counterfactuals might not only depend on a cortical hierarchy of sufficient size to support an inner loop separable from immediate engagement with the sensorium (Buckner and Krienen, 2013), but also a working H/E-S to stabilize ensembles associated with novel (due to being non-actual) possibilities (Hassabis et al., 2007). In this way, in conjunction with the rest of the cortex, the H/E-S could be viewed as an energy-based self-supervised contrastive learner (Mazzaglia et al., 2022), which may enable a substantial amount of adaptive-autonomous behavior if (variational) free-energy/prediction-error is being minimized with respect to divergences between goals and present estimated states (Hafner et al., 2020; Safron, 2021b).

It has recently been suggested by researchers at Numenta (a biologically-inspired AI company) that the principles (and particular cellular adaptations such as grid cells) involved in H/E-S functioning—e.g., allocentric object modeling (Sabour et al., 2017; Kosiorek et al., 2019)—may be recapitulated throughout the entire neocortex (Hawkins et al., 2019). The idea that the H/E-S may represent a template for understanding the neocortex is not unreasonable, since while it is referred to as "subcortical", it is technically composed of cortical tissue (Insausti et al., 2017). Along these lines, not only is the H/E-S topologically central as a "convergence divergence zone" (Damasio, 2012) and hub for "semantic pointers" (Blouw et al., 2016), but it is also primary from an evolutionary (as archaecortex/periallocortex) and developmental perspective.

Modeling based on object-centered reference frames may be a broader property of the neocortex (Hawkins, 2021). However, we believe that such coherent perspectives may depend on being able to conduct active inference and learning with sufficient degrees of independence from other modeling/control processes (Thomas et al., 2017, 2018). That is, we suggest that for emergent modules to have H/E-S properties, they must be able to achieve informational closure with sufficient rapidity that they can both independently inform and be informed by action-perception cycles with respect to particular effector-sensor systems. For example, the establishment of such independently controllable factors may be the case for large macrocolumns such as rodent whisker barrels, but potentially not for ocular dominance columns. To the extent that hippocampal and entorhinal cell-types are found more generally throughout the cortex (Long and Zhang, 2021), we suggest that it remains ambiguous as to whether this reflects G-SLAM constituting a common cortical algorithm, or whether such representations are induced over the course of development *via* integrative functioning involving the H/E-S.

The H/E-S as sense-maker and value integrator/realizer

Switching between conceptual scenes involves ramping of hippocampal activity, followed by high-frequency signaling with the cortex as a new frame of sense making is established (de la Prida, 2020; Karimi Abadchi et al., 2020; O'Callaghan et al., 2021). Theoretically, these events (potentially accompanied by sharp wave ripples) would represent the formation of new grid/place tiling/mapping/graphing over a space/scope of relevance, but where sufficient functionality is carried over across remappings for integration of information across episodes. Functionally speaking, these frame-shifts could be understood in terms of Lévy flights with respect to generalized

search, so allowing for more exploratory processing and creative solutions in the face of challenges (Hills et al., 2010; McNamee et al., 2021). That is, in contrast to searching *via* random walks that would tend to result in reliable exploitative mapping of simple domains, such discontinuous (and potentially fanciful) flights to remote areas of hypothesis/phase spaces would allow agents to both more efficiently explore complex domains and escape from local optima. Considering that the H/E-S may be understood as the highest (or most flexibly integrative) level of agent-level control processes, altering parameters/modulators relevant for this kind of more exploitative or exploratory (generalized) search could be some of the most significant sources on variation both between and within individuals and species (Safron, 2020c).

While the precise conditions for remapping are likely to vary based on multiple conditions, degree of overall prediction-error seems to be one reliable trigger, as in an experiment in which participants were cued to retrieve well-learned complex room images from memory and then presented with either identical or modified pictures (Bein et al., 2020); in this study, the number of changes caused CA1–CA3 connectivity to decrease (potentially indicating less intra-hippocampal recurrent activity) and CA1-entorhinal connectivity to increase. Consistently, another study found sensitivity to reward prediction errors with respect to the establishment of new event boundaries (Rouhani et al., 2020). Similar influences on the stability of mappings by more general salience is suggested by studies in which the H/E-S shows sensitivity to interactions with the amygdala and responses to fearful stimuli (Chen et al., 2019), as well as modulation of encoding based on attention/expectancy (Mack et al., 2018, 2020; Urgolites et al., 2020). The dividing of continuous unfoldings into discrete epochs provides another means by which abstract phenomena such as time may be conceptualized by the H/E-S (in addition to their spatialization, perhaps as a kind of multidimensional scaling onto lower dimensional manifolds that may be inspected either through fictive navigation or imaginative visual foraging (Ramachandran et al., 2016).

Notably, the H/E-S may not just be sensitive to reward, but it may also help to provide a major source of the prediction errors that drive phasic dopamine (Mannella et al., 2013; Ballard et al., 2019; Jang et al., 2019; Laubach et al., 2020), potentially involving internal contrasting between hippocampal subfields, and with overall prediction-error being further integrated *via* outputs to the subiculum (Tingley and Buzsáki, 2018; Canto et al., 2019). This may allow for the allostatic prioritization of goals with respect to not only cortical predictions from medial prefrontal cortices, but even homeostatic regulatory nuclei of the septum (Tingley and Buzsáki, 2018; Kunz et al., 2019; Livneh et al., 2020). The importance of the H/E-S for motivational states is also evidenced by its ability to influence the interoceptive components of emotions (Edwards-Duric et al., 2020), which may have a further (circular) causal significance in helping to drive counterfactual simulations, potentially understandable as affectively-canalized Markov chain Monte Carlo tree search through value space (Dohmatob et al., 2020; Hesp et al., 2020; Parascandolo et al., 2020; Safron, 2021b). In this way, not only would the H/E-S help implement SLAM processes with respect to both concrete and abstract cognition, but it may also help to explain how agent-level mental processes can enter causal streams leading to both mental simulations and overt enaction, so affording some of the varieties of "free will" worth having for autonomous systems (Safron, 2021b).

Some evidence for this affective influencing of H/E-S dynamics may potentially be found in studies of increased inter-hemispheric phase coupling (delta range coherence) during treadmill running periods (Furtunato et al., 2020), potentially corresponding to periods of increased driving by biophysical signals indicating organismic salience. Crucially, sources of H/E-S "reward" may not just take the form of the aforementioned extrinsic value of goal realization, but may also be driven by the intrinsic value of novel information, for the hippocampus could provide a natural integrator of prediction-error as top of the cortical hierarchy (Hawkins and Blakeslee, 2004; Mannella et al., 2013; Fonken et al., 2020). While the hippocampus and ventromedial prefrontal cortex may usually work together in estimating expected value (or opportunities for free energy minimization), theoretically, they may also function as semi-separate value signals in terms of respective information gain and preference satisfaction. In this way, convergence of the H/E-S and its ventromedial prefrontal collaborators upon the accumbens core—and thereby nigral motor dopamine (Mannella et al., 2013)—may represent physical manifestations of the dual optimization for intrinsic and extrinsic value prescribed by active inference as a normative account of intelligence. This kind of convergent control based on heterogeneous (fundamental) value signals is notable, as it is becoming increasingly clear the H/E-S is more than just a temporary memory buffer, but rather may constitute a primary basis for autonomous functioning for vertebrates as adaptive cybernetic systems, as highlighted in **Figure 6**.

G(eneralized-)SLAM as core cognitive process

As described above, the H/E-S and its functional relationships with the neocortex may be understood as implementing a kind of Kalman variational autoencoder (Fraccaro et al., 2017). In this capacity, the H/E-S may provide inspiration for developing advanced SLAM architectures. In its dual role as both memory and control system, the H/E-S has been further optimized for facilitating comparisons between largescale patterns (e.g., organismic states), which in machine learning terms may be understood as implementing something akin to energy-based contrastive learning

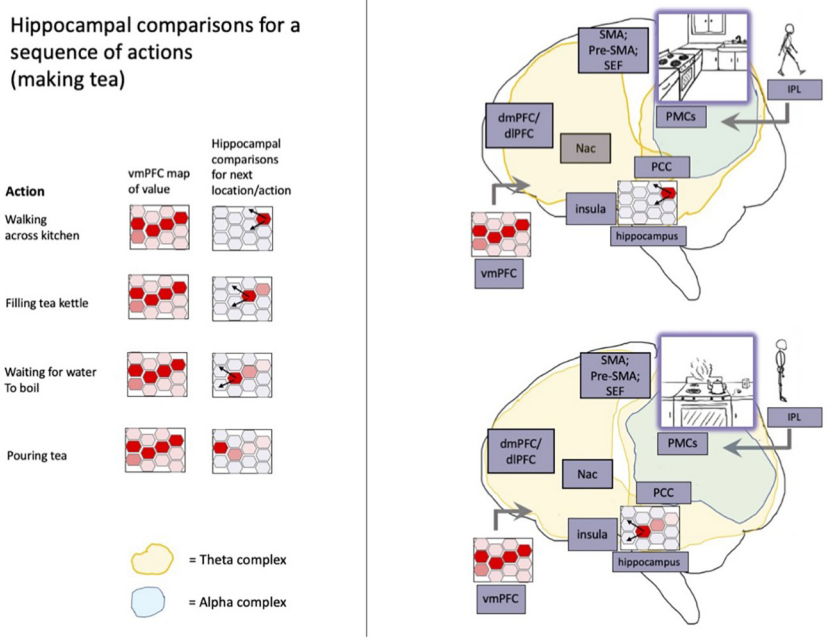


FIGURE 6
A model of hippocampally-orchestrated imaginative planning and action selection *via* generalized navigation. Goal-oriented action sequences are depicted with respect to relevant neural processes. The hippocampal system provides (a) organization of cortical attracting states into value-canalized spatiotemporal trajectories, (b) stabilization of ensembles *via* theta-mediated cross-frequency phase coupling, and (c) goal-oriented cognition and behavior *via* contrasting (not depicted) sensed and imagined states. Hippocampal trajectories are shaped according to whichever paths are expected to result in more positively valanced outcomes (cf. reward prediction errors). The expected value associated with navigating to different portions of (potentially abstract) space is informed *via* coupling with similarly spatiotemporally-organized value representations (red shaded hexagons) in vmPFC and associated systems. As chained patterns of activity progress across hippocampal place fields (red hexagons with variable degrees of shading), theta-synchronized frontal ensembles (yellow shading spreading towards the front of the brain) help to generate (*via* cross-frequency phase coupling) ensembles for directing attention, working memory, and overt enaction. Sensory updating of posterior cortices occurs at alpha frequencies (blue shading), so providing a basis for conscious perception and imagination. With respect to these integrated estimates of sensory states, hippocampal coupling at theta frequencies (yellow shading spreading towards the back of the brain) provides a basis for (a) episodic memory and replay, (b) novel imaginings, and (c) adjustment of neuronal activity selection *via* orchestrated contrasting between cortical ensembles. Abbreviations: nAC, nucleus accumbens; vmPFC, ventromedial prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; SMA, supplementary motor area; Pre-SMA, presupplementary motor area; SEF, supplementary eye fields; PCC, posterior cingulate cortex; PMCs, posterior medial cortices; IPL, inferior parietal lobule. Reprinted with permission from Safron (2021b).

(Marblestone et al., 2016; Richards et al., 2019; Mazzaglia et al., 2022). In this capacity, the H/E-S may provide inspiration for developing architectures capable of engaging in self-supervised learning, counterfactual modeling, and further enabling high-level reasoning abilities including analogical structure mapping (Gentner, 2010; Safron, 2019a), causal inference (Pearl and Mackenzie, 2018), and imaginative planning (Kaplan and Friston, 2018; Safron, 2021b).

As described above, and elsewhere (Safron, 2020b,c, 2021a,b), LatentSLAM's dual-tier architecture provides an abstract cybernetic interpretation of the H/E-S as the highest (or most integrative) level of heterarchical control for embodied-embedded organisms as they move through physical and imagined worlds in the pursuit of valued goals, so providing a computational/functional account of agency in biological (and perhaps artificial) systems. Further, this hierarchical architecture provides a basis for meta-learning in which slower and more encompassing "outer loop" processes aggregate information over faster "inner loop" processes, so affording the much-desired goal of realizing synergistic inference and generalization of knowledge across experiences (or lessons in curriculums for lifelong learning). Even more, the upper levels of this kind of hybrid architecture may provide a basis for explicit symbolic reasoning (*via* abstract experience graphs) in addition to enactive couplings with the world (*via* adaptive control of poses/views), both of which are likely required for achieving the goal of robust autonomous functioning for artificial systems.

While attempting to navigate towards such destinations may seem excessively ambitious, we would note that work on the extended H/E-S was part of what inspired the formation of some of the world's leading AI companies such as DeepMind,

and continues to be a central part of their research programs (Hassabis and Maguire, 2009; Hassabis et al., 2017; Koster et al., 2018; McNamee et al., 2021). Indeed, it is increasingly being recognized that the spatiotemporal modeling properties of the H/E-S may constitute an invaluable integrative framework for understanding high-level cognition (Whittington et al., 2022). However, we believe a G-SLAM framing might be particularly notable in connecting to the context under which these systems were first selected/shaped by evolution (and development), as well as one of the primary functionalities of the H/E-S that continues throughout the lifespan of organisms. That is, our abilities to navigate both physical and conceptual worlds represent an ongoing challenge for as long as we live. We further suggest the connection between the practical necessities involved in engineering physical systems may provide a particularly valuable source of empirical traction for attempting to specify the roles of particular features of the H/E-S, in that we can draw upon the rich data generated as robots attempt to navigate through the world.

Further, by also drawing upon biological details in designing AI-architectures, we may find ourselves with access to invaluable inductive biases which might be otherwise overlooked. Two examples that come to mind include recent proposals by Bengio and LeCun with respect to "GFlowNets" and "Joint Embedding Predictive Architectures" (Bengio et al., 2022; LeCun, 2022). We believe these efforts in creating autonomous and generally intelligent systems may benefit by incorporating principles of G-SLAM, such as the creation of systems capable of handling loop trajectories as potentially enabling greater open-ended and life-long learning, or in looking towards hybrid systems similar to LatentSLAM as potentially allowing for explicit representations and symbolic processing. While it has often been said that the goal of AI is to create the "cognitive equivalent of an airplane wing," we would suggest that the magnitude of the challenge may be far greater (more akin to building a fully functioning plane or space ship), and the problems of navigating through under-constrained architectural (and learning curricula) design-spaces may be unsurmountable without biological inspiration/grounding.

While LatentSLAM continues to be refined, we believe these kinds of architectures provide a general framework for understanding core elements of minds and brains. Indeed, to localize something within a spatialized reference frame—which itself is impacted by the entities it maps/graphs— may be what it means to "understand" and "explain" something (Lakoff and Johnson, 1999), and possibly even to experience anything at all (Safron, 2020a, 2021a,b). That is to understand is to be able to adopt a stance (or pose) from which elements and their inter-relations may be mapped (or localized), as if projected onto a plane whereby they are made visible for inspection (or navigation). We believe these etymological considerations on the nature of knowledge may be more than "mere" metaphors but could point to the fundamentally embodied nature of minds.

We not only suggest that all thought may be understood as navigating between representations that are being localized and mapped (or graphed) within an organizing conceptual domain, but all communication may be understood as the transmission of such structures (as trajectories) between minds (Zurn and Bassett, 2020). While the simultaneity of generalized localization and mapping in cognition may not be obvious upon introspection, this is more clearly the case when considering unfamiliar concepts. For such novel domains, relationships between concepts and broader organizing schemas involves the same kind of challenges of circular inference as found in SLAM. That is, when we are first attempting to understand a conceptual domain, we do not know how to effectively connect the entities whose shared features and relations motivate the construction of organizing schemas. However, without such higher-order abstractions and the predictive (or compressive) capacities they provide, it is unclear which features of and relations between entities are relevant for shared structure learning.

Heuristic algorithms may be invaluable in the bootstrapping process, such as the kind of clustering involved in the hierarchical Dirichlet process (Griffiths T. et al., 2019), models of category formation *via* analogical alignment (Kuehne et al., 2000), or concept derivation as abstraction over episodes (Mack et al., 2018). We agree that such accounts may speak to fundamental mental processes, but we also suggest that rather than static feature maps, such nonparametric (Bayesian) structure learning may apply to paths through mapped/graphed domains. This is part of why we emphasized our use of the Fisher distance measure above, as an information metric that naturally applies to trajectories may potentially provide the most valid (and potentially predictive) means of assessing similarity/dissimilarity between entities in feature spaces. Indeed, one of the most notable aspects of thinking is its sequential operation and sensitivity to path dependencies. While abstract conceptualization does allow for a good deal of cognitive flexibility, cognition is still largely defined by deriving knowledge *via* particular "chains of reasoning," or "paths" through mental space.

Regardless of the particular routes by which we reach the heights of category learning, the formation of such abstract representations constitute what may be the most powerful aspect of our intelligence in terms of generalizable knowledge that can robustly transfer across particular episodes (Marcus, 2020). Such abstract categories further allow for the kinds of structured representations whose importance was emphasized in decades of work in (non-radically-enactive) cognitive science and "good old fashioned AI." The significance of such knowledge structures may prove even greater in light of the advent of graph networks within the context of geometric deep learning (Battaglia et al., 2018)—and symbolic regression as potentially representing a further degree of abstraction (Cranmer et al., 2020). Such graphical representations are of increasing interest because of both their interpretability as well as their extraordinary

efficiency for modeling physical systems. With our models of node duplication and graph-relaxation, LatentSLAM provides a biologically plausible and computationally-tractable account of how such cognitive schemas may be formed and modified through experience. This is notable in that finding principled means of creating and modifying particular structures for graph neural networks (GNNs) remains an ongoing challenge. But if such challenges can be surmounted, then we may achieve the promise of neurosymbolic AI in combining the power of connectionism with reasoning over explicitly represented (and related) symbols (Garcez and Lamb, 2020; Greff et al., 2020). More specifically, we believe that the ability of the H/E-S to create navigable spaces populated by high-level attracting states may also provide a basis for creating *"ad hoc"* (Barsalou, 1983) GNN structures for different purposes.

While the relationships between place cells in the H/E-S (or nodes in LatentSLAM) can be understood as a kind of GNN, we believe it would be more accurate to characterize these models as graph nets, in that they represent relations—or semantic pointers (Blouw et al., 2016)—for hierarchically lower graphs. While these details have yet to be incorporated into LatentSLAM, it has been suggested that heteromodal association cortices may constitute a shared latent space across (autoencoding) cortical hierarchies with quasi-topographic characteristics akin to those found with GNNs (Safron, 2020b, 2021a,b). While the H/E-S has significant interactions with the entire cortical heterarchy, connectivity is most substantial for deeper (or hierarchically higher) portions of cortex, consistent with its potential role as a kind of graph network. The degree to which these machine learning analogies may apply to brain functioning is yet to be determined, but they nonetheless represent a promising direction for creating artificial systems that recapitulate the properties of natural intelligences (Greff et al., 2020).

Intriguingly, the work in which brains were proposed to entail GNN-type computation was developed independently of LatentSLAM. However, similarly to how LatentSLAMs only uses views and proprioceptive poses for specifying particular experiences to be mapped (**Figure 2**), this other work proposed that sufficient bases for agentic world modeling may involve conjoined visuospatial and somatospatial modalities, potentially (but not necessarily) understood as respective grid and mesh-pose GNNs. In the model of episodic memory and imagination described above (**Figure 6**), H/E-S trajectories are used to orchestrate state transitions between these experiences as the "stream of consciousness" (James, 1890). While many aspects of cognition are unconscious, "thinking" and "reasoning" are usually considered to involve sequentially generated conscious operations. Notably, the formal conceptualization of computation may have been largely inspired by Turing introspecting his own consciousness in the process of doing mathematics (Dehaene, 2014; Graves et al., 2014). Given that it is unclear that we can be conscious of anything that lacks grounding in somatic modalities and their abilities to change (and be controlled) through time, then all thinking/reasoning may potentially be understood as involving the kinds of action selection and modeling described by LatentSLAM.

Fully describing the potential correspondences between SLAM and high-level cognition is beyond the scope of a single publication (**Table 1**), but before concluding we will briefly comment on the importance of loop-closures and thresholds for graph-relaxation and node duplication. In brief, we may understand a (generalized) loop-closure event as a primary factor contributing to the feeling of understanding and insight (Gopnik, 1998; Fonken et al., 2020; Oh et al., 2020). After an initial period of relatively ambiguous exploration, the formation of a causal account (or trajectory through a concept space/graph) would allow for a rapid decrease in prediction-error (Joffily and Coricelli, 2013), or increase in compression (Schmidhuber, 2010). While some individuals may be relatively insensitive to these feelings of (potentially sudden) conceptual familiarity (Hou et al., 2013; Ben-Yakov et al., 2014), others may potentially be overly sensitive (e.g., "déjà vu" and other kinds of false positive inferences), with the specific functional tradeoffs involved depending on particular contexts (DeYoung, 2015; Blain et al., 2020; Safron and DeYoung, 2021).

Events in which this kind of cognitive closure is achieved provide special opportunities for updating H/E-S models (or categories) *via* graph-relaxation and node duplication. A variety of relevant parameters can be identified (**Figure 3**), whether in terms of thresholds for detecting loop-closures, the extent to which graphs may be relaxed, or the ease with which new nodes are created. However, in a G-SLAM context of trying to model cognition more generally, we may think of loop-closure recognition thresholds as sensitivity to cumulative prediction error increases/decreases, graph-relaxation as changing attractor dynamics within the H/E-S and neocortex on multiple scales, and node duplication as the establishment of new local ensembles of effective connectivity (cf. chained bump attractors)—and potentially (but not necessarily) involving neurogenesis, for which it is notable that the hippocampus is one of the few places where this phenomenon is reliably observed. These core SLAM processes may depend on multiple factors, including neuromodulators such as dopamine and serotonin (Safron, 2020c; Safron and Sheikhbahaee, 2021), as well as on Bayesian priors (or "yesterday's posteriors").

With respect to the previously described example of differential tuning thresholds for mapping the structure of aisles (**Figure 5**), we may potentially have a crucial source of individual differences in cognition. In theory, G-SLAM may be pointing to (or localizing) a cognitive spectrum (and potential basis for differential diatheses) spanning autism and schizophrenia (Byars et al., 2014; Crespi and Dinsdale, 2019). Theoretically, we may even expect to see these kinds of variations in SLAM maps in the drawings of autistic and schizophrenic individuals (Morgan et al., 2019; Philippsen and Nagai, 2020).

Speculatively, not only may the conceptual understanding of that which is being drawn be mapped and navigated by the H/E-S as SLAM system, but the eye movements (Wynn et al., 2020) and hand motions involved in skilled actions such as drawing could themselves be orchestrated according to hippocampal trajectories as a basis for chained equilibrium setpoints (Latash, 2010). Even more speculatively, it could even be the case that further degrees of sophisticated control—as inference (Kaplan and Friston, 2018; Friston et al., 2021)—are bootstrapped by simultaneously localizing and mapping the body itself as a kind of space/graph, so allowing for more rarefied and general SLAM capacities over the course of development. In this view, much of cognitive development would involve initial phases of using the H/E-S to learn intentional control over either overtly or covertly expressed motor patterns, which then become automatized (or amortized) by the thalamic-cerebellar system (Safron, 2021a; Shine, 2021) and dorsal striatal-cortical loops (Mannella et al., 2013), so freeing up the G-SLAM system for further high-level predictive modeling and control. Can the body itself be understood as a mapped spatial domain, or is this just a way of speaking without any useful technical correspondences? How far can we go with using these patterns of linguistic use as hypotheses regarding cognitive processes and underlying neural mechanisms? Could it even be the case that the phenomenology of embodiment involves navigation through and mapping of body maps *via* these cross-modal interactions, which when disrupted could potentially contribute to altered states of consciousness or potentially clinical conditions such as depersonalization (Safron, 2020c; Ciaunica and Safron, 2022)?

With respect to personhood, beyond its foundational role for autonomous functioning, widespread orchestration of value-canalized trajectories through biophysical phase space by the H/E-S also enables the development (and ongoing functioning) of the spatiotemporally-extended processes required for autonoetic and autobiographical self-consciousness. In addition to constituting major transitions in evolution, the advent of such self-reflective capacities may have been required for the construction of advanced social coordination and a (shared) symbolic order of being. While such rarefied processes may be well-beyond anything we are close to engendering in (abiotic) machines, it may be the case that we are forced to recapitulate these kinds of H/E-S functionalities if we are to successfully arrive at the destination of creating robustly autonomous and general artificial intelligences.

Indeed, G-SLAM parameters may constitute the most important source of variation we can identify both between and within individuals. To venture deep into unknown speculative territory, the H/E-S may be the source of key adaptations contributing to the evolution of cognitive modernity through (potentially proto-schizotypal) flexibly creative cognition and the birth of cumulative culture, which in time came to represent what may be the "secret of our success" as a species and the greatest of all major transitions in (generalized) evolution (Premack, 1983; Gentner, 2010; Hofstadter and Sander, 2013; Henrich, 2017; Safron, 2019b, 2020c; van den Heuvel et al., 2019; Dehaene et al., 2022). While such models extend far beyond domains of knowledge for which we have well-developed maps, we believe such possibilities are worthy of further exploration.

Present limitations and future directions for G-SLAM

While we describe experiments for LatentSLAM in other publications (Çatal et al., 2021a), future work should attempt to explicitly illustrate G-SLAM principles with experiments and mathematical models/simulations. Further, while approaches to localization and mapping may be diverse, this does not mean that all technical solutions involved are best described as SLAM problems. However, we believe that analogues of processes like loop closure and node duplication (and pruning) with respect to trajectories through cognitive spaces would constitute strong evidence for the value of a generalized SLAM perspective. It is also important to note that symbolic processing in the brain involves more than the H/E-S. For instance, a substantial amount of symbolic communication is linguistic in a way that could be described in terms of a hierarchical control system for vocal production and hearing (gestural communication could provide another illustrative example). While such action-perception cycles need not involve the H/E-S, we also believe their functioning may potentially be enhanced *via* H/E-S orchestration of high-level dynamics (e.g., channeling neuronal manifolds along particular trajectories).

We also believe it will be valuable to explore research attempting to combine SLAM and various forms of semantic processing in robotics/AI (Kostavelis and Gasteratos, 2015; Sünderhauf et al., 2017; Garg et al., 2020). Not only does such work illustrate the complexity of SLAM problems and how they may (and must) be integrated with other cognitive processes (cf. artificial consciousness?), but it also points to other ways in which robotics can be used to inform our understanding of minds, whether biologically grown and artificially engineered). Finally, while we focus on a particular SLAM architecture developed within the Free Energy Principle and Active Inference framework, we believe it will be fruitful to consider other approaches as well, many of which are extremely well developed and sophisticated in their own right (Penny et al., 2013; Madl et al., 2018; Stoianov et al., 2022; Taniguchi et al., 2022).

Conclusions

We have searched through broad and diverse terrains in considering the ideas above, covering a lot of ground. To try

to come full circle, we have described technical details of a machine learning architecture for autonomous robot navigation, discussed particulars of biological systems for realizing these functionalities in brains, and started to explore how these principles may provide a framework for understanding all high-level cognition in terms of simultaneous localization and mapping in space (broadly construed to include conceptual spaces). We have only begun this journey, but we believe the destination is promising, and we invite others to join us in exploring this framework for understanding the nature of thought. Some might contend that "prediction" or "modeling" are more encompassing and fundamental than a generalized SLAM perspective, and we would not disagree. However, we believe that G-SLAM is unique in allowing for all these perspectives to be combined with the principles of ecological rationality that constituted the primary selective pressures for high-level cognition over the course of evolution and development. We suggest this neuroethological perspective will be invaluable in allowing us to "carve nature at its joints", in terms of identifying the most important features of functioning for the hippocampal/entorhinal system and its connections to the rest of the brain [and body (and world)]. We further believe that G-SLAM is unique in the extent to which it connects to nature(s) of experience, where we do in fact exist in a spatial world through which we must navigate, and where it is difficult to find aspects of mind not impacted by this fundamental physical situatedness. In light of these sources of potential insight, we believe that G-SLAM represents the way forward for understanding complex minds, and potentially for building them, if we can find sustainable paths into the unexplored territory of the future.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Author contributions

AS, OÇ, and TV conceptualized G-SLAM and the main ideas for this manuscript. OÇ and TV conceived and performed the LatentSLAM experiments. AS contributed his knowledge gained from his ongoing study of the hippocampal/entorhinal system literature. All authors contributed to the article and approved the submitted version.

Funding

OÇ is funded by a Ph.D. grant of the Flanders Research Foundation (FWO). This research received funding from the "AI Flanders" program of the Flemish government.

Acknowledgments

We would like to express thanks to many people for many reasons. Hopefully you know who you are. We hope to share more of our gratitude with you in time (and space).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ai, H., Okada, R., Sakura, M., Wachtler, T., and Ikeno, H. (2019). Neuroethology of the waggle dance: how followers interact with the waggle dancer and detect spatial information. *Insects* 10:336. doi: 10.3390/insects10100336

Ambrose, R. E., Pfeiffer, B. E., and Foster, D. J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron* 91, 1124–1136. doi: 10.1016/j.neuron.2016.07.047

Andersen, P. (1975). "Organization of hippocampal neurons and their interconnections," in *The Hippocampus: Volume 1: Structure and Development*, eds R. L. Isaacson, and K. H. Pribram (Boston, MA: Springer US), 155–175.

Ball, D., Heath, S., Wiles, J., Wyeth, G., Corke, P., and Milford, M. (2013). OpenRatSLAM: an open source brain-based SLAM system. *Auton. Robots* 34, 149–176. doi: 10.1007/s10514-012-9317-9

Ballard, I. C., Wagner, A. D., and McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* 10:1073. doi: 10.1038/s41467-019-08998-1

Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Prog. Neurobiol.* 192:101821. doi: 10.1016/j.pneurobio.2020.101821

Barsalou, L. W. (1983). Ad hoc categories. *Mem. Cognit.* 11, 211–227. doi: 10.3758/BF03196968

Bastos, A. M., Lundqvist, M., Waite, A. S., Kopell, N., and Miller, E. K. (2020). Layer and rhythm specificity for predictive routing. *Proc. Natl. Acad. Sci.* 117, 31459–31469. doi: 10.1073/pnas.2014868117

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning and graph networks. *arXiv* [Preprint]. doi: 10.48550/arXiv.1806.01261

Bein, O., Duncan, K., and Davachi, L. (2020). Mnemonic prediction errors bias hippocampal states. *Nat. Commun.* 11:3451. doi: 10.1038/s41467-020-17287-1

Bellmund, J. L. S., Cothi, W. de., Ruiter, T. A., Nau, M., Barry, C., and Doeller, C. F. (2019). Deforming the metric of cognitive maps distorts memory. *Nat. Hum. Behav.* 1–12. doi: 10.1038/s41562-019-0767-3

Bengio, Y. (2017). The consciousness prior. *arXiv* [Preprint]. doi: 10.48550/arXiv.1709.08568

Bengio, Y., Deleu, T., Hu, E. J., Lahlou, S., Tiwari, M., and Bengio, E. (2022). GFlowNet Foundations. *arXiv* [Preprint]. doi: 10.48550/arXiv.2111.09266

Ben-Yakov, A., Rubinson, M., and Dudai, Y. (2014). Shifting gears in hippocampus: temporal dissociation between familiarity and novelty signatures in a single event. *J. Neurosci.* 34, 12973–12981. doi: 10.1523/JNEUROSCI.1892-14.2014

Bergen, B. K. (2012). *Louder Than Words: The New Science of How the Mind Makes Meaning*. New York, NY: Basic Books.

Birch, J., Ginsburg, S., and Jablonka, E. (2020). Unlimited associative learning and the origins of consciousness: a primer and some predictions. *Biol. Philos.* 35:56. doi: 10.1007/s10539-020-09772-0

Blain, S. D., Longenecker, J. M., Grazioplene, R. G., Klimes-Dougan, B., and DeYoung, C. G. (2020). Apophenia as the disposition to false positives: a unifying framework for openness and psychoticism. *J. Abnorm. Psychol.* 129, 279–292. doi: 10.1037/abn0000504

Blouw, P., Solodkin, E., Thagard, P., and Eliasmith, C. (2016). Concepts as semantic pointers: a framework and computational model. *Cogn. Sci.* 40, 1128–1162. doi: 10.1111/cogs.12265

Boccara, C. N., Nardin, M., Stella, F., O'Neill, J., and Csicsvari, J. (2019). The entorhinal cognitive map is attracted to goals. *Science* 363, 1443–1447. doi: 10.1126/science.aav4837

Buckner, R. L., and Krienen, F. M. (2013). The evolution of distributed association networks in the human brain. *Trends Cogn. Sci.* 17, 648–665. doi: 10.1016/j.tics.2013.09.017

Butler, W. N., Hardcastle, K., and Giocomo, L. M. (2019). Remembered reward locations restructure entorhinal spatial maps. *Science* 363, 1447–1452. doi: 10.1126/science.aav5297

Byars, S. G., Stearns, S. C., and Boomsma, J. J. (2014). Opposite risk patterns for autism and schizophrenia are associated with normal variation in birth size: phenotypic support for hypothesized diametric gene-dosage effects. *Proc. R. Soc. B Biol. Sci.* 281:20140604. doi: 10.1098/rspb.2014.0604

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., et al. (2016). Past, present and future of simultaneous localization and mapping: toward the robust-perception age. *Trans. Robot.* 32, 1309–1332. doi: 10.1109/TRO.2016.2624754

Canolty, R. T., and Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends Cogn. Sci.* 14, 506–515. doi: 10.1016/j.tics.2010.09.001

Canto, C. B., Koganezawa, N., Lagartos, M. J. D., O'Reilly, K. C., Mansvelder, H. D., and Witter, M. P. (2019). Postnatal development of functional projections from para- and presubiculum to medial entorhinal cortex in the rat. *J. Neurosci.* 1, 1623–19. doi: 10.1523/JNEUROSCI.1623-19.2019

Çatal, O., Jansen, W., Verbelen, T., Dhoedt, B., and Steckel, J. (2021a). "LatentSLAM: unsupervised multi-sensor representation learning for localization and mapping," in *2021 International Conference on Robotics and Automation (ICRA)* (Xi'an, China), 6739–6745. doi: 10.1109/ICRA48506.2021.9560768

Çatal, O., Verbelen, T., Maele, T. V. de., Dhoedt, B., and Safron, A. (2021b). Robot navigation as hierarchical active inference. *Neural Netw.* 142, 192–204. doi: 10.1016/j.neunet.2021.05.010

Chen, B. K., Murawski, N. J., Cincotta, C., McKissick, O., Finkelstein, A., Hamidi, A. B., et al. (2019). Artificially enhancing and suppressing hippocampus-mediated memories. *Curr. Biol.* 29, 1885–1894. doi: 10.1016/j.cub.2019.04.065

Ciaunica, A., and Safron, A. (2022). Disintegrating and reintegrating the self - (In)flexible self-models in depersonalisation and psychedelic experiences. *PsyArxiv* [Preprint]. doi: 10.31234/osf.io/mah78

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Costa, S. I. R., Santos, S. A., and Strapasson, J. E. (2015). Fisher information distance: a geometrical reading. *Discrete Appl. Math.* 197, 59–69. doi: 10.1016/j.dam.2014.10.004

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., et al. (2020). Discovering symbolic models from deep learning with inductive biases. *arXiv* [Preprint]. doi: 10.48550/arXiv.2006.11287

Crespi, B., and Dinsdale, N. (2019). Autism and psychosis as diametrical disorders of embodiment. *Evol. Med. Public Health* 2019, 121–138. doi: 10.1093/emph/eoz021

Crouse, M., Nakos, C., Abdelaziz, I., and Forbus, K. (2020). Neural analogical matching. *arXiv* [Preprint]. doi: 10.48550/arXiv.2004.03573

Damasio, A. (2012). *Self Comes to Mind: Constructing the Conscious Brain*. New York: Vintage.

Dasgupta, I., Schulz, E., Goodman, N. D., and Gershman, S. J. (2018). Remembrance of inferences past: amortization in human hypothesis generation. *Cognition* 178, 67–81. doi: 10.1016/j.cognition.2018.04.017

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., et al. (2019). Causal reasoning from meta-reinforcement learning. *arXiv* [Preprint]. doi: 10.48550/arXiv.1901.08162

de la Prida, L. M. (2020). Potential factors influencing replay across CA1 during sharp-wave ripples. *Philos. Trans. R. Soc. B Biol. Sci.* 375:20190236. doi: 10.1098/rstb.2019.0236

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Viking.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., and Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends Cogn. Sci.* 26, 751–766. doi: 10.1016/j.tics.2022.06.010

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W. W. Norton and Company.

DeYoung, C. G. (2015). Cybernetic big five theory. *J. Res. Personal.* 56, 33–58. doi: 10.1016/j.jrp.2014.07.004

Dohmatob, E., Dumas, G., and Bzdok, D. (2020). Dark control: the default mode network as a reinforcement learning agent. *Hum. Brain Mapp.* 41, 3318–3341. doi: 10.1002/hbm.25019

Edwards-Duric, J., Stevenson, R. J., and Francis, H. M. (2020). The congruence of interoceptive predictions and hippocampal-related memory. *Biol. Psychol.* 155:107929. doi: 10.1016/j.biopsycho.2020.107929

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* 15, 732–744. doi: 10.1038/nrn3827

Eichenbaum, H. (2015). The hippocampus as a cognitive map . . . of social space. *Neuron* 87, 9–11. doi: 10.1016/j.neuron.2015.06.013

Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P. D. H., Butz, M. V., and Wermter, S. (2020). Hierarchical principles of embodied reinforcement learning: a review. *arXiv* [Preprint]. doi: 10.48550/arXiv.2012.10147

Evans, T., and Burgess, N. (2020). Replay as structural inference in the hippocampal-entorhinal system. *bioRxiv* [Preprint]. doi: 10.1101/2020.08.07.241547

Feinberg, T. E., and Mallatt, J. (2013). The evolutionary and genetic origins of consciousness in the Cambrian period over 500 million years ago. *Front. Psychol.* 4:667. doi: 10.3389/fpsyg.2013.00667

Fonken, Y. M., Kam, J. W. Y., and Knight, R. T. (2020). A differential role for human hippocampus in novelty and contextual processing: implications for P300. *Psychophysiology* 57:e13400. doi: 10.1111/psyp.13400

Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. (2017). A disentangled recognition and nonlinear dynamics model for unsupervised learning. *arXiv* [Preprint]. doi: 10.48550/arXiv.1710.05741

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K. J., Parr, T., and de Vries, B. (2017b). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K., Da Costa, L., Hafner, D., Hesp, C., and Parr, T. (2021). Sophisticated inference. *Neural Comput.* 33, 713–763. doi: 10.1162/neco_a_01351

Friston, K., Parr, T., and Zeidman, P. (2019). Bayesian model reduction. *arXiv* [Preprint]. doi: 10.48550/arXiv.1805.07092

Furtunato, A. M. B., Lobão-Soares, B., Tort, A. B. L., and Belchior, H. (2020). Specific increase of hippocampal delta oscillations across consecutive treadmill runs. *Front. Behav. Neurosci.* 14:101. doi: 10.3389/fnbeh.2020.00101

Garcez, A. d. A., and Lamb, L. C. (2020). Neurosymbolic AI: the 3rd wave. *arXiv* [Preprint]. doi: 10.48550/arXiv.2012.05876

Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., et al. (2020). Semantics for robotic mapping, perception and interaction: a survey. *Found. Trends Robot.* 8, 1–224. doi: 10.1561/2300000059

Gauthier, B., Pestke, K., and van Wassenhove, V. (2019). Building the arrow of time. Over time: a sequence of brain activity mapping imagined events in time and space. *Cereb. Cortex* 29, 4398–4414. doi: 10.1093/cercor/bhy320

Gentner, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cogn. Sci.* 34, 752–775. doi: 10.1111/j.1551-6709.2010.01114.x

George, D., Rikhye, R. V., Gothoskar, N., Guntupalli, J. S., Dedieu, A., and Lázaro-Gredilla, M. (2021). Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* 12:2392. doi: 10.1038/s41467-021-22559-5

Gershman, S., and Goodman, N. (2014). Amortized inference in probabilistic reasoning. *Proc. Annu. Meet. Cogn. Sci. Soc.* 36. Available online at: https://escholarship.org/uc/item/34j1h7k5. Accessed August 26, 2020.

Gopnik, A. (1998). Explanation as orgasm. *Minds Mach.* 8, 101–118. doi: 10.1023/A:1008290415597

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv* [Preprint]. doi: 10.48550/arXiv.1410.5401

Gray, J. A., and McNaughton, N. (2003). *The Neuropsychology of Anxiety: An Enquiry Into the Function of the Septo-Hippocampal System*. Oxford: Oxford University Press.

Greff, K., van Steenkiste, S., and Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv* [Preprint]. doi: 10.48550/arXiv.2012.05208

Griffiths, B. J., Parish, G., Roux, F., Michelmann, S., Plas, M. v. d., Kolibius, L. D., et al (2019). Directional coupling of slow and fast hippocampal gamma with neocortical alpha/beta oscillations in human episodic memory. *Proc. Natl. Acad. Sci.* 116, 21834–21842. doi: 10.1073/pnas.1914180116

Griffiths, T., Canini, K., Sanborn, A., and Navarro, D. (2019). Unifying rational models of categorization via the hierarchical dirichlet process. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/ketw3

Hafner, D., Ortega, P. A., Ba, J., Parr, T., Friston, K., and Heess, N. (2020). Action and perception as divergence minimization. *arXiv* [Preprint]. doi: 10.48550/arXiv.2009.01791

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806. doi: 10.1038/nature03721

Hassabis, D., and Maguire, E. A. (2009). The construction system of the brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1263–1271. doi: 10.1098/rstb.2008.0296

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hassabis, D., Kumaran, D., Vann, S. D., and Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci.* 104, 1726–1731. doi: 10.1073/pnas.0610561104

Hawkins, J. (2021). *A Thousand Brains: A New Theory of Intelligence*. New York: Basic Books.

Hawkins, J., and Blakeslee, S. (2004). *On Intelligence Adapted*. New York: Times Books.

Hawkins, J., Lewis, M., Klukas, M., Purdy, S., and Ahmad, S. (2019). A framework for intelligence and cortical function based on grid cells in the neocortex. *Front. Neural Circuits* 12:121. doi: 10.3389/fncir.2018.00121

Henrich, J. (2017). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Hesp, C., Tschantz, A., Millidge, B., Ramstead, M., Friston, K., and Smith, R. (2020). "Sophisticated affective inference: simulating anticipatory affective dynamics of imagining future events," in *Active Inference Communications in Computer and Information Science*, eds. T. Verbelen, P. Lanillos, C. L. Buckley, and C. De Boom (Cham: Springer International Publishing), 179–186.

Hess, W., Kohler, D., Rapp, H., and Andor, D. (2016). "Real-time loop closure in 2D LIDAR SLAM," in *2016 IEEE International Conference on Robotics and Automation (ICRA)* (Stockholm, Sweden), 1271–1278. doi: 10.1109/ICRA.2016.7487258

Heusser, A. C., Poeppel, D., Ezzyat, Y., and Davachi, L. (2016). Episodic sequence memory is supported by a theta-gamma phase code. *Nat. Neurosci.* 19, 1374–1380. doi: 10.1038/nn.4374

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., et al. (2020). Replay bursts coincide with activation of the default mode and parietal alpha network. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.23.166645

Hills, T. T., Kalff, C., and Wiener, J. M. (2013). Adaptive lévy processes and area-restricted search in human foraging. *PLoS One* 8:e60488. doi: 10.1371/journal.pone.0060488

Hills, T. T., Todd, P. M., and Goldstone, R. L. (2010). The central executive as a search process: priming exploration and exploitation across domains. *J. Exp. Psychol. Gen.* 139, 590–609. doi: 10.1037/a0020666

Hofstadter, D., and Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York: Basic Books.

Honkanen, A., Adden, A., Freitas, J. da S., and Heinze, S. (2019). The insect central complex and the neural basis of navigational strategies. *J. Exp. Biol.* 222:jeb188854. doi: 10.1242/jeb.188854

Hou, M., Safron, A., Paller, K. A., and Guo, C. (2013). Neural correlates of familiarity and conceptual fluency in a recognition test with ancient pictographic characters. *Brain Res.* 1518, 48–60. doi: 10.1016/j.brainres.2013.04.041

Howard, M. W. (2018). Memory as perception of the past: compressed time in mind and brain. *Trends Cogn. Sci.* 22, 124–136. doi: 10.1016/j.tics.2017.11.004

Igata, H., Ikegaya, Y., and Sasaki, T. (2020). Prioritized experience replays on a hippocampal predictive map for learning. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.23.002964

Insausti, R., Muñoz-López, M., Insausti, A. M., and Artacho-Pérula, E. (2017). The human periallocortex: layer pattern in presubiculum, parasubiculum and entorhinal cortex. A review. *Front. Neuroanat.* 11:84.doi: 10.3389/fnana.2017.00084

James, W. (1890). *The Principles of Psychology, Vol. 1*. New York: Dover Publications.

Jang, A. I., Nassar, M. R., Dillon, D. G., and Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nat. Hum. Behav.* 3, 719–732. doi: 10.1038/s41562-019-0597-3

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Kalman, R. E., and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *J. Basic Eng.* 83, 95–108. doi: 10.1115/1.3658902

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Karimi Abadchi, J., Nazari-Ahangarkolaee, M., Gattas, S., Bermudez-Contreras, E., Luczak, A., McNaughton, B. L., et al. (2020). Spatiotemporal patterns of neocortical activity around hippocampal sharp-wave ripples. *eLife* 9:e51972. doi: 10.7554/eLife.51972

Kirsch, L., and Schmidhuber, J. (2020). Meta learning backpropagation and improving it. *arXiv* [Preprint]. doi: 10.48550/arXiv.2012.14905

Kosiorek, A., Sabour, S., Teh, Y. W., and Hinton, G. E. (2019). "Stacked capsule autoencoders," in *Advances in Neural Information Processing Systems 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, R. Garnett, et al (Red Hook, NY: Curran Associates, Inc.), 15433. Available online at: https://www.proceedings.com/content/053/053719webtoc.pdf.

Kostavelis, I., and Gasteratos, A. (2015). Semantic mapping for mobile robotics tasks: a survey. *Robot. Auton. Syst.* 66, 86–103. doi: 10.1016/j.robot.2014.12.006

Koster, R., Chadwick, M. J., Chen, Y., Berron, D., Banino, A., Düzel, E., et al. (2018). Big-loop recurrence within the hippocampal system supports integration of information across episodes. *Neuron* 99, 1342–1354. doi: 10.1016/j.neuron.2018.08.009

Kuehne, S. E., Forbus, K. D., Gentner, D., and Quinn, B. (2000). SEQL: category learning as progressive abstraction using structure mapping. Available online at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.4757.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Kunz, L., Wang, L., Lachner-Piza, D., Zhang, H., Brandt, A., Dümpelmann, M., et al. (2019). Hippocampal theta phases organize the reactivation of large-scale electrophysiological representations during goal-directed navigation. *Sci. Adv.* 5:eaav8192. doi: 10.1126/sciadv.aav8192

Lakoff, G., and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

Latash, M. L. (2010). Motor synergies and the equilibrium-point hypothesis. *Motor Control* 14, 294–322. doi: 10.1123/mcj.14.3.294

Laubach, M., Amarante, L. M., Caetano, M. S., and Horst, N. K. (2020). Reward signaling by the rodent medial frontal cortex. *Int. Rev. Neurobiol.* 158, 115–133. doi: 10.1016/bs.irn.2020.11.012

LeCun, Y. (2022). A path towards autonomous machine intelligence. *OpenReview* Available online at: https://openreview.net/forum?id=BZ5a1r-kVsf. Accessed June 28, 2022.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., and Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* 29, 9771–9777. doi: 10.1523/JNEUROSCI.1319-09.2009

Liashenko, A., Dizaji, A. S., Melloni, L., and Schwiedrzik, C. M. (2020). Memory guidance of value-based decision making at an abstract level of representation. *Sci. Rep.* 10:21496. doi: 10.1038/s41598-020-78460-6

Livneh, Y., Sugden, A. U., Madara, J. C., Essner, R. A., Flores, V. I., Sugden, L. A., et al. (2020). Estimation of current and future physiological states in insular cortex. *Neuron* 105, 1094–1111. doi: 10.1016/j.neuron.2019.12.027

Long, X., and Zhang, S.-J. (2021). A novel somatosensory spatial navigation system outside the hippocampal formation. *Cell Res.* 31, 649–663. doi: 10.1038/s41422-020-00448-8

Maass, A., Schütze, H., Speck, O., Yonelinas, A., Tempelmann, C., Heinze, H.-J., et al. (2014). Laminar activity in the hippocampus and entorhinal cortex related to novelty and episodic encoding. *Nat. Commun.* 5, 1–12. doi: 10.1038/ncomms6547

Mack, M. L., Love, B. C., and Preston, A. R. (2018). Building concepts one episode at a time: the hippocampus and concept formation. *Neurosci. Lett.* 680, 31–38. doi: 10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., and Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* 11:46. doi: 10.1038/s41467-019-13930-8

MacKay, D. G. (2019). *Remembering: What 50 Years of Research with Famous Amnesia Patient H. M. Can Teach Us about Memory and How It Works*. Amherst, MA: Prometheus Books.

Madl, T., Franklin, S., Chen, K., and Trappl, R. (2018). A computational cognitive framework of spatial memory in brains and robots. *Cogn. Syst. Res.* 47, 147–172. doi: 10.1016/j.cogsys.2017.08.002

Mannella, F., Gurney, K., and Baldassarre, G. (2013). The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Front. Behav. Neurosci.* 7:135. doi: 10.3389/fnbeh.2013.00135

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10:94. doi: 10.3389/fncom.2016.00094

Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. *arXiv* [Preprint]. doi: 10.48550/arXiv.2002.06177

Mazzaglia, P., Verbelen, T., and Dhoedt, B. (2022). Contrastive active inference. *arXiv* [Preprint]. doi: 10.48550/arXiv.2110.10083

McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2021). Flexible modulation of sequence generation in the entorhinal-hippocampal system. *Nat. Neurosci.* 24, 851–862. doi: 10.1038/s41593-021-00831-7

Milford, M. J., Wyeth, G. F., and Prasser, D. (2004). "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *2004 International Conference on Robotics and Automation (ICRA)* (New Orleans, LA), 403–408. doi: 10.1109/ROBOT.2004.1307183

Mok, R. M., and Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nat. Commun.* 10:5685. doi: 10.1038/s41467-019-13760-8

Mok, R. M., and Love, B. C. (2020). Abstract neural representations of category membership beyond information coding stimulus or response. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.13.947341

Morgan, A. T., Petro, L. S., and Muckli, L. (2019). Line drawings reveal the structure of internal visual models conveyed by cortical feedback. *bioRxiv* [Preprint]. doi: 10.1101/041186

Mulders, D., Yim, M. Y., Lee, J. S., Lee, A. K., Taillefumier, T., and Fiete, I. R. (2021). A structured scaffold underlies activity in the hippocampus. *bioRxiv* [Preprint]. doi: 10.1101/2021.11.20.469406

Mur-Artal, R., Montiel, J. M. M., and Tardós, J. D. (2015). ORB-SLAM: a Versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* 31, 1147–1163. doi: 10.1109/TRO.2015.2463671

O'Callaghan, C., Walpola, I. C., and Shine, J. M. (2021). Neuromodulation of the mind-wandering brain state: the interaction between neuromodulatory tone, sharp wave-ripples and spontaneous thought. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 376:20190699. doi: 10.1098/rstb.2019.0699

O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Oh, Y., Chesebrough, C., Erickson, B., Zhang, F., and Kounios, J. (2020). An insight-related neural reward signal. *Neuroimage* 214:116757. doi: 10.1016/j.neuroimage.2020.116757

Papez, J.W. (1937). A proposed mechanism of emotion. *Arch. Neurol. Psychiatry* 38, 725–743.

Parascandolo, G., Buesing, L., Merel, J., Hasenclever, L., Aslanides, J., Hamrick, J. B., et al. (2020). Divide-and-conquer monte carlo tree search for goal-directed planning. *arXiv* [Preprint]. doi: 10.48550/arXiv.2004.11410

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322–1327. doi: 10.1126/science.1159775

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Penny, W. D., Zeidman, P., and Burgess, N. (2013). Forward and backward inference in spatial cognition. *PLoS Comput. Biol.* 9:e1003383. doi: 10.1371/journal.pcbi.1003383

Philippsen, A., and Nagai, Y. (2020). A predictive coding account for cognition in human children and chimpanzees: a case study of drawing. *IEEE Trans. Cogn. Dev. Syst.* 1:1. doi: 10.1109/TCDS.2020.3006497

Premack, D. (1983). The codes of man and beasts. *Behav. Brain Sci.* 6, 125–136. doi: 10.1017/S0140525X00015077

Quiroga, R. Q. (2020). No pattern separation in the human hippocampus. *Trends Cogn. Sci.* 24, 994–1007. doi: 10.1016/j.tics.2020.09.012

Ramachandran, V. S., Vajanaphanich, M., and Chunharas, C. (2016). Calendars in the brain; their perceptual characteristics and possible neural substrate. *Neurocase* 22, 461–465. doi: 10.1080/13554794.2016.1250913

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Rouhani, N., Norman, K. A., Niv, Y., and Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition* 203:104269. doi: 10.1016/j.cognition.2020.104269

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv* [Preprint]. doi: 10.48550/arXiv.1710.09829

Safron, A. (2019a). Bayesian analogical cybernetics. *arXiv* [Preprint]. doi: 10.48550/arXiv.1911.02362

Safron, A. (2019b). Multilevel evolutionary developmental optimization (MEDO): a theoretical framework for understanding preferences and selection dynamics. *arXiv* [Preprint]. doi: 10.48550/arXiv.1910.13443

Safron, A. (2020a). An integrated world modeling theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Front. Artif. Intell.* 3:30. doi: 10.3389/frai.2020.00030

Safron, A. (2020b). Integrated world modeling theory (IWMT) implemented: towards reverse engineering consciousness with the free energy principle and active inference. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/paz5j

Safron, A. (2020c). On the varieties of conscious experiences: altered beliefs under psychedelics (ALBUS). *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/zqh4b

Safron, A. (2021a). Integrated world modeling theory (IWMT) expanded: implications for theories of consciousness and artificial intelligence. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/rm5b2

Safron, A. (2021b). The radically embodied conscious cybernetic bayesian brain: from free energy to free will and back again. *Entropy* 23:783. doi: 10.3390/e23060783

Safron, A., and DeYoung, C. G. (2021). "Chapter 18 - integrating cybernetic big five theory with the free energy principle: a new strategy for modeling personalities as complex systems," in *Measuring and Modeling Persons and Situations*, eds. D. Wood, S. J. Read, P. D. Harms, and A. Slaughter (London, UK: Academic Press), 617–649.

Safron, A., and Sheikhbahaee, Z. (2021). Dream to explore: 5-HT2a as adaptive temperature parameter for sophisticated affective inference. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/zmpaq

Schmidhuber, J. (2010). Formal theory of creativity, fun and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368

Schmidhuber, J. (2020). Reinforcement learning upside down: don't predict rewards – just map them to actions. *arXiv* [Preprint]. doi: 10.48550/arXiv.1912.02875

Shamash, P., Olesen, S. F., Iordanidou, P., Campagner, D., Nabhojit, B., Branco, T., et al. (2020). Mice learn multi-step routes by memorizing subgoal locations. *bioRxiv* [Preprint]. 2020.08.19.256867. doi: 10.1101/2020.08.19.256867

Shang, W., Trott, A., Zheng, S., Xiong, C., and Socher, R. (2019). Learning world graphs to accelerate hierarchical reinforcement learning. *arXiv* [Preprint]. doi: 10.48550/arXiv.1907.00664

Sharif, F., Tayebi, B., Buzsáki, G., Royer, S., and Fernandez-Ruiz, A. (2020). Subcircuits of deep and superficial CA1 place cells support efficient spatial coding across heterogeneous environments. *Neuron* 109, 363–376. doi: 10.1016/j.neuron.2020.10.034

Shine, J. M. (2021). The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. *Prog. Neurobiol.* 199:101951. doi: 10.1016/j.pneurobio.2020.101951

Spiers, H. J., Hayman, R. M. A., Jovalekic, A., Marozzi, E., and Jeffery, K. J. (2015). Place field repetition and purely local remapping in a multicompartment environment. *Cereb. Cortex* 25, 10–25. doi: 10.1093/cercor/bht198

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* 20, 1643–1653. doi: 10.1038/nn.4650

Stoianov, I., Maisto, D., and Pezzulo, G. (2022). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *Prog. Neurobiol.* 217:102329. doi: 10.1016/j.pneurobio.2022.102329

Striedter, G. F. (2004). *Principles of Brain Evolution*. Sunderland, MA: Sinauer Associates is an imprint of Oxford University Press.

Sünderhauf, N., Pham, T. T., Latif, Y., Milford, M., and Reid, I. (2017). "Meaningful maps with object-oriented semantic mapping," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC, Canada), 5079–5085. doi: 10.1109/IROS.2017.8206392

Suryanarayana, S. M., Pérez-Fernández, J., Robertson, B., and Grillner, S. (2020). The evolutionary origin of visual and somatosensory representation in the vertebrate pallium. *Nat. Ecol. Evol.* 1–13. doi: 10.1038/s41559-020-1137-2

Taniguchi, A., Fukawa, A., and Yamakawa, H. (2022). Hippocampal formation-inspired probabilistic generative model. *Neural Netw.* 151, 317–335. doi: 10.1016/j.neunet.2022.04.001

Thomas, V., Bengio, E., Fedus, W., Pondard, J., Beaudoin, P., Larochelle, H., et al. (2018). Disentangling the independently controllable factors of variation by interacting with the world. *arXiv* [Preprint]. doi: 10.48550/arXiv.1802.09484

Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., et al. (2017). Independently controllable factors. *arXiv* [Preprint]. doi: 10.48550/arXiv.1708.01289

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Thrun, S., and Montemerlo, M. (2006). The graph SLAM algorithm with applications to large-scale mapping of urban structures. *Int. J. Robot. Res.* 25, 403–429. doi: 10.1177/0278364906065387

Tingley, D., and Buzsáki, G. (2018). Transformation of a spatial map across the hippocampal-lateral septal circuit. *Neuron* 98, 1229–1242. doi: 10.1016/j.neuron.2018.04.028

Todd, P. M., and Gigerenzer, G. (2012). *Ecological Rationality: Intelligence in the World*. Oxford; New York: Oxford University Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208. doi: 10.1037/h0061626

Tversky, B. (2019). *Mind in Motion: How Action Shapes Thought*. New York: Basic Books.

Umbach, G., Kantak, P., Jacobs, J., Kahana, M., Pfeiffer, B. E., Sperling, M., et al. (2020). Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proc. Natl. Acad. Sci.* 117, 28463–28474. doi: 10.1073/pnas.2013250117

Urgolites, Z. J., Wixted, J. T., Goldinger, S. D., Papesh, M. H., Treiman, D. M., Squire, L. R., et al. (2020). Spiking activity in the human hippocampus prior to encoding predicts subsequent memory. *Proc. Natl. Acad. Sci.* 117, 13767–13770. doi: 10.1073/pnas.2001338117

Uria, B., Ibarz, B., Banino, A., Zambaldi, V., Kumaran, D., Hassabis, D., et al. (2020). The spatial memory pipeline: a model of egocentric to allocentric understanding in mammalian brains. *bioRxiv* [Preprint]. doi: 10.1101/2020.11.11.378141

van den Heuvel, M. P., Scholtens, L. H., de Lange, S. C., Pijnenburg, R., Cahn, W., van Haren, N. E. M., et al. (2019). Evolutionary modifications in human brain connectivity associated with schizophrenia. *Brain J. Neurol.* 142, 3991–4002..doi: 10.1093/brain/awz330

Wang, C., Chen, X., Lee, H., Deshmukh, S. S., Yoganarasimha, D., Savelli, F., et al. (2018). Egocentric coding of external items in the lateral entorhinal cortex. *Science* 362, 945–949. doi: 10.1126/science.aau4940

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* 21:860. doi: 10.1038/s41593-018-0147-8

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., and Behrens, T. E. J. (2022). How to build a cognitive map: insights from models of the hippocampal formation. *arXiv* [Preprint]. doi: 10.48550/arXiv.2202.01682

Whittington, J. C. R., Muller, T. H., Mark, S., Barry, C., and Behrens, T. E. J. (2018). Generalisation of structural knowledge in the hippocampal-entorhinal system. *arXiv* [Preprint]. doi: 10.48550/arXiv.1805.09042

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., et al. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* 183, 1249–1263. e23. doi: 10.1016/j.cell.2020.10.024

Wijesinghe, R., Protti, D. A., and Camp, A. J. (2015). Vestibular Interactions in the Thalamus. *Front. Neural Circuits* 9:79. doi: 10.3389/fncir.2015.00079

Wynn, J. S., Ryan, J. D., and Buchsbaum, B. R. (2020). Eye movements support behavioral pattern completion. *Proc. Natl. Acad. Sci.* 117, 6246–6254. doi: 10.1073/pnas.1917586117

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* 10, 1–7. doi: 10.1038/s41467-019-11786-6

Zhang, F., Li, S., Yuan, S., Sun, E., and Zhao, L. (2017). "Algorithms analysis of mobile robot SLAM based on Kalman and particle filter," in *2017 9th International Conference on Modelling, Identification and Control (ICMIC)* (Kunming, China), 1050–1055. doi: 10.1109/ICMIC.2017.8321612

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., et al. (2019). Graph neural networks: a review of methods and applications. *arXiv* [Preprint]. doi: 10.48550/arXiv.1812.08434

Zurn, P., and Bassett, D. S. (2020). Network architectures supporting learnability. *Philos. Trans. R. Soc. B Biol. Sci.* 375:20190323. doi: 10.1098/rstb.2019.0323

Appendix A: LatentSLAM mathematical model

The latentSLAM model for SLAM can be seen as a two-level active inference hierarchy working together to enable navigation. The lower-level abstracts actions and observations from the physical world into an abstract representation. The higher-level takes the lower-level abstractions as inputs and creates a global abstraction over them. In this appendix we will go into the mathematical details of these models.

Generic active inference model

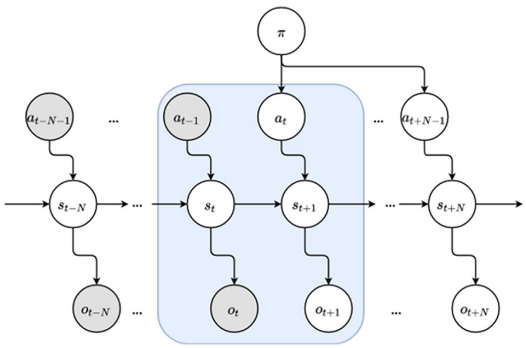


FIGURE A1
Graphical representation of a generic active inference model. Gray nodes indicate that the corresponding value is observed, white nodes need to be inferred.

Both levels of the hierarchy form an instantiation of an active inference model (Figure A1). This means that each level forms a generative model over its own actions and observations. We assume the environment is modeled up to a certain time horizon T by the agent as a POMDP with joint probability distribution

$$P(\tilde{o}, \tilde{s}, \tilde{a}, \pi) = P(s_o)P(\pi)\prod_{t=1}^{T} P(o_t|s_t)P(s_t|s_{t-1}, a_{t-1})P(a_{t-1}|\pi)$$

Where tildes indicate sequences of the corresponding variables, a indicates the action, o the observation, s the latent states and π the policy.

The agent needs to infer the posterior belief on latent states $P(\tilde{s}|\tilde{o}, \tilde{a})$. In order to achieve this, we use a variational approximation of the true posterior, which we parametrize as

$$Q(\tilde{s}|\tilde{o}, \tilde{a}) = Q(s_0|o_0)\prod_{t=1}^{T} Q(s_t|s_{t-1}, a_{t-1}, o_t)$$

Note that in all following discussions we will use Q to designate a (variational) posterior and P as a prior distribution.

As we the agent is acting according to the free energy principle, it is actively minimizing its variational free energy. Which we posit here as

$$\mathrm{F} = D_{KL}[Q(\tilde{s}|\tilde{o}, \tilde{a})||P(\tilde{s}, \tilde{a})] - \mathbb{E}_{Q(\tilde{s}|\tilde{o},\tilde{a})}[\log P(\tilde{o}|\tilde{s})]$$

For a more detailed description of the derivation of F, we refer the reader to. The generative model and the free energy form only one aspect of active inference. The agent not only needs to infer states from the present, but also actions for the future. This is achieved through the expected free energy, which we define for a future timestep τ and a given policy π as

$$G(\pi, \tau) = D_{KL}[Q(s_\tau|\pi)||P(s_\tau)] + \mathbb{E}_{Q(s_\tau)}[H(P(o_\tau|s_\tau))]$$

Summing over the future timesteps then gives the expected free energy for a given policy.

Navigation model

Similar to the generic active inference model, we start by defining the generative model in terms of a joint distribution over its parts.

$$P(\tilde{o}, \tilde{s}, \tilde{a}, \tilde{l}, \tilde{p}, \tilde{m}, \pi) = P(\tilde{o}, \tilde{s}_{i>0}, \tilde{a}, \tilde{p}_{i>0}, \pi|\tilde{l}, \tilde{s}_0, \tilde{p}_0)P(\tilde{l}, \tilde{m}, \tilde{s}_0, \tilde{p}_0)$$

Where o, s, and π keep their earlier definition and we now introduce the pose p, location l and move m to the discussion. This joint distribution naturally decomposes into two new joints over a subset of variables, allowing the independent treatment of higher-level and lower-level navigation. The resulting graphical model is shown in Figure A2.

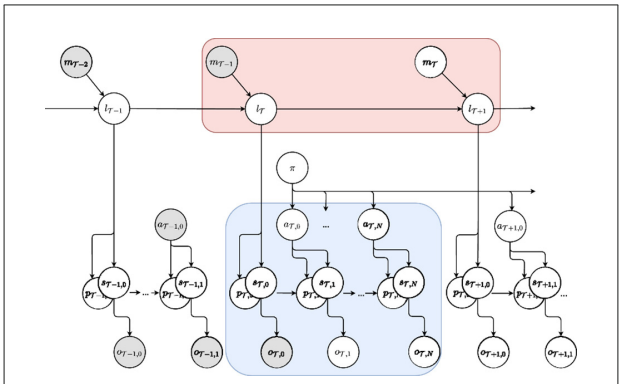


FIGURE A2
Graphical overview of the hierarchical navigation model.

If we look at the free energy of this model, we get

$$\begin{aligned} F_{hierarchical} &= \mathbb{E}_Q\left[\log Q\left(\tilde{s},\tilde{p}\right) - \log P\left(\tilde{o},\tilde{s}_{i>0},\tilde{a},\tilde{p}_{i>0}|\tilde{l},\tilde{s}_0,\tilde{p}_o\right)\right. \\ &\quad \left. +\log Q\left(\tilde{l}\right) - \log P\left(\tilde{l},\tilde{m},\tilde{s}_0,\tilde{p}_0\right)\right] \\ &= F_{low} + F_{high} \end{aligned}$$

Allowing for a clean separation in the creation of the lower and higher-level state models.

Lower-level model

Using the same approximations for the lower-level model as in the generic active inference model, we write the free energy as

$$\begin{aligned} F_{low} &= \sum_t \mathbb{E}_Q[\log Q(p_t) - \log P(p_t|p_{t-1},a_{t-1},s_t)] \\ &\quad + D_{KL}[Q(s_t)||P(s_t|s_{t-1},a_{t-1})] + \mathbb{E}_Q[-\log P(o_t|s_t)] \end{aligned}$$

From this we recover the same generative model for the observations and observational latent states as before in the generic case, however, the model is now supplemented with a term responsible for the pose estimation aspects. In effect, this means that we do not use any pose information for the visual perception part of the model. Note that the pose estimation is in fact conditioned on the current perceptual state estimate.

As might be expected from the free energy formulation, the lower-level perception is implemented as a generic active inference model. The pose estimation aspect is deliberately left as an expected difference between the pose posterior and prior.

The pose prior is implemented as simple dynamics model from the action velocities used to control the agent, i.e.,

$$\begin{aligned} \theta_t &= \theta_{t-1} + v_a\Delta_t \\ x_t &= x_{t-1} + v_l\Delta_t \\ y_t &= y_{t-1} + v_l\Delta_t \end{aligned}$$

With $v_a$and v_l the angular and linear velocity of the agent and x,y,θ the coordinates and rotation in the plane of the agent.

The pose posterior $Q\ (p_t)$ is implemented as a CAN with energy dynamics described as

$$\begin{aligned} \varepsilon_{\Delta x,\Delta y,\Delta\theta} &= \exp\frac{-\Delta x^2-\Delta y^2}{k_p^{exc}}\exp\frac{-\Delta\theta^2}{k_d^{exc}} \\ &\quad - \exp\frac{-\Delta x^2-\Delta y^2}{k_d^{inh}}\exp\frac{-\Delta\theta^2}{k_d^{inh}} \end{aligned}$$

With k_d and k_p the variance constants for place and direction, and the superscript *exc* and *inh* used to indicate whether the effect is inhibitory or excitatory. The resulting behavior is locally excitatory and globally inhibitory. The conditioning on observatory state s_t is achieved by creating an extra excitatory link with a state-pose episodic memory.

Higher-level model

Starting again from the free energy functional

$$\begin{aligned} F_{high} &= \sum_{\mathcal{T}} D_{KL}[Q(l_{\mathcal{T}})||P(l_{\mathcal{T}}|l_{\mathcal{T}-1},m_{\mathcal{T}-1})] \\ &\quad + \mathbb{E}_Q[-\log P(s_{\mathcal{T},0}|l_{\mathcal{T}}) - \log P(p_{\mathcal{T},0}|l_{\mathcal{T}})] \end{aligned}$$

We again see the classical active inference model emerging. Note the usage of $\mathcal{T}$ instead of t to indicate that this model operates on a different timescale. Likewise, only the initial lower-level states for that inference cycle appear in the likelihood (remember that this model uses the states of the lower-level as observations). Here again, the actual implementation of these models is geared towards a navigational task. In order to infer the location, the prior distribution of locations is implemented as an experience graph. Each node in the graph incorporates a state, pose pair to link it with the lower level. Links between the nodes indicate a connection traversable on the lower level. The dynamics model $P(l_{\mathcal{T}}|l_{\mathcal{T}-1},m_{\mathcal{T}}-1)$ is deduced from the adjacency matrix of the graph. The posterior belief $Q(l_{\mathcal{T}}|s_{\mathcal{T},t},p_{\mathcal{T},t})$ is build by assigning probability inversely proportional to the cosine similarity and Euclidean distance between current state pose pair and the pairs reachable from the current node.

Map updates also trigger a graph-relaxation pass, in order to facilitate loop-closures. The graph-relaxation phase shifts the stored poses in each experience map node according to

$$\Delta p^i = \frac{1}{2}\left[\sum_{j=1}^{inbound}(p^j - p^i - \Delta p^{ij}) + \sum_{k=1}^{outbound}(p^k - p^i - \Delta p^{ij})\right]$$

Action inference

So far, we have only discussed the state inference aspects of the navigation model, however, action inference is also an important aspect of active inference. The expected free energy for the lower-level model is

$$\begin{aligned} G_{low}(\pi,\tau) &= D_{KL}[Q(s_\tau,p_\tau|\pi)||Q(s_{T+1},p_{T+1}|l_T,m_T)] \\ &\quad + \mathbb{E}_{Q(s_\tau)}[H(P(o_\tau|s_\tau))] \end{aligned}$$

The prior preferences in this equation are provided by the higher-level model, and form targets to achieve within the single

timestep of that level. High level targets are then extracted according to

$$G_{high}(\pi, \tau) = D_{KL}[Q(l_\tau|\pi)||P(l_\tau)] + E_{Q(l_\tau)}[H(P(p_{\tau,0}|l_\tau)) + H(P(s_{\tau,0}|l_\tau))]$$

Action selection happens than according to a two-phase planning process. First, at the coarser higher-level, second at the fine-grained lower level. This allows for a reduction in search space for a given trajectory.

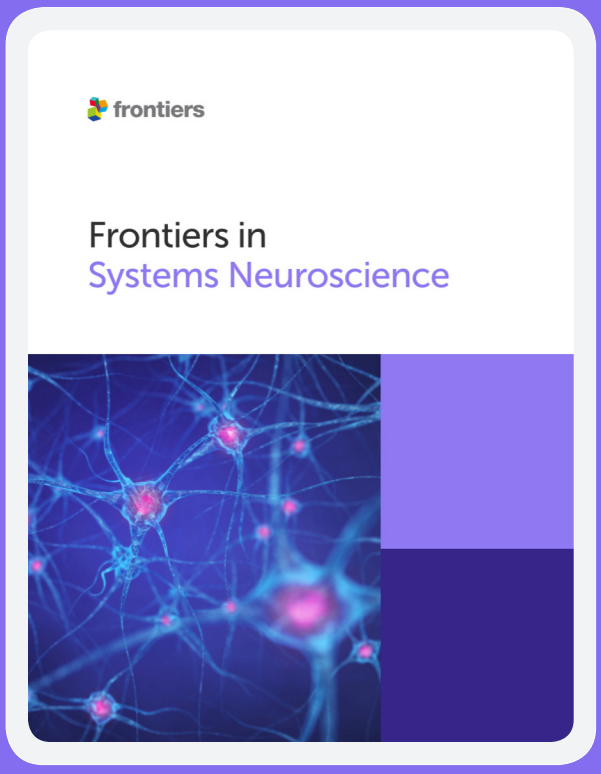
Frontiers in
Systems Neuroscience

Advances our understanding of whole systems of the brain

Part of the most cited neuroscience journal series, this journal explores the architecture of brain systems and information processing, storage and retrieval.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact