frontiers | Frontiers in Computer Science

TYPE Hypothesis and Theory
PUBLISHED 18 November 2025
DOI 10.3389/fcomp.2025.1662185

OPEN ACCESS

EDITED BY
Pavlos Papadopoulos,
Edinburgh Napier University, United Kingdom

REVIEWED BY
Dimitrios Kasimatis,
Edinburgh Napier University, United Kingdom
Binny Jose,
Marian College Kuttikkanam Autonomous, India
Alexander Stevens,
Memorial Sloan Kettering Cancer Center, United State

*CORRESPONDENCE
Gene Michael Alarcon
✉ gene.alarcon.1@us.af.mil

RECEIVED 14 July 2025
ACCEPTED 29 September 2025
PUBLISHED 18 November 2025

CITATION
Alarcon GM and Capiola A (2025) Explicating the trust process for effective human interaction with artificial intelligence and machine learning systems.
Front. Comput. Sci. 7:1662185.
doi: 10.3389/fcomp.2025.1662185

Explicating the trust process for effective human interaction with artificial intelligence and machine learning systems

Gene Michael Alarcon* and August Capiola

Air Force Research Laboratory, Dayton, OH, United States

Artificial intelligence (AI) and machine learning (ML) are rapidly changing the landscape of almost every environment. Despite the burgeoning attention on this subject matter, limited human-centered research has focused on understanding how users interact with AI and ML to facilitate greater trust toward these systems, leveraging classic human-machine interaction principles to investigate human interaction with these emerging complex systems. The current paper incorporates literature from Social Psychology, Computer Science, Information Sciences, and Human Factors Psychology to create a single comprehensive model for understanding user interactions with AI/ML-enabled systems. This paper expands previous theoretical models by explicating transparency, incorporating individual differences in the information processing model of cognition, and summarizing the different attitudes and personality variables that can facilitate use and disuse of AI and ML. The theoretical model proposed explicitly demarcates the referent algorithm from the human user, detailing the processes that eventuate a user's reliance on and compliance with an AI/ML-enabled system. Actual and potential applications of the literature review and theorized model are discussed.

KEYWORDS

trust, human-machine interaction, artificial intelligence, machine learning, explainable artificial intelligence

Introduction

Artificial intelligence (AI) and machine learning (ML) are rapidly changing the technological landscape of work. AI/ML have been utilized in a variety of tasks such as image classification (e.g., Hendrycks et al., 2019), route planning (e.g., Hu et al., 2020), and parole recommendations (e.g., Hübner, 2021), to name a few (see Kaplan et al., 2020). These applications of AI/ML have brought increased productivity and alleviated aspects of work that were previously reserved for human oversight and assumed to be outside the realm of machine capability. However, the applications of AI/ML have also brought questions as to how to facilitate proper trust in and use of these algorithms.

The current paper seeks to build a theoretical model of human-AI/ML interaction, incorporating the underlying information that both AI/ML algorithms and explainable AI (XAI)[1] tools use as input and provide as output, and explicate the psychological processes that lead to trust in AI/ML/XAI from an information processing theory perspective. In this

1 We note we use the term AI/ML/XAI when referring to all models and their respective acronyms for other models.

paper we leverage research from many fields including Computer Science, Psychology, and Information Sciences. The current paper seeks to add clarity to the field to help guide research on interface design(s) relevant to human-AI/ML/XAI interaction, with the goal of increasing appropriate use of AI/ML/XAI-enabled systems. Importantly, our proposed model can be applied to all human-AI/ML/XAI interactions. Our model expands information processing theory (Wickens and Hollands, 2000; Wickens, 2002) by situating individual differences from the Organizational, Personality, and Social Psychology literatures as meaningful constructs that may shape users' processing of information from AI/ML/XAI referents. We also expand on the different attitudes that can result from this information processing. Lastly, our theoretical model also expands the theory on machine transparency, explaining how individual differences influence the cognitive processing of the outputs of the AI/ML/XAI information. The goal of the paper is to incorporate all previous knowledge into one workable framework that can be utilized across disciplines.

AI, ML, and XAI

AI/ML

AI/ML are advanced algorithms that enable software systems with human-like cognitive capacity for decision-making (Dwivedi et al., 2021; Herm et al., 2021; Hradecky et al., 2022). AI/ML are powerful generalizers and predictors (Burrell, 2016) that have demonstrated their use in a variety of settings (Barreto Arrieta et al., 2020). The application of AI/ML to prediction cases is not new. It is both the rapid increase of AI/ML into a variety of fields and the opaqueness of many of the emerging algorithms that are relatively new (Ali et al., 2023). Newer algorithms utilize complex networks of nodes and hidden layers that predict outcomes. These nodes and hidden layers are much more complex than traditional algorithms like regression. The increased complexity leads to a lack of comprehension as to how the model came to a decision, either locally (e.g., a given instance) or globally (e.g., how the model works overall; see Zhou et al., 2021).

At its simplest form, the Computer Science literature has largely demarcated algorithms into white- and black-box models (Herm et al., 2023).[2] White-box models are relatively less complex AI/ML-enabled algorithms, where the underlying decision processes are understandable to the user (e.g., regression, decision trees, and generalized additive models). Their processes for reaching a given classification, and the variables that are or are not important to reach a given prediction, are comparatively interpretable by human users (Barreto Arrieta et al., 2020). White-box models typically require feature selection of pertinent characteristics for the development of an algorithm. In contrast, black-box models utilize all relevant data and interactions in non-linear models to predict and generalize. Black-box models are often so complex they provide little to no information about the underlying decision process to the user (Lundberg, 2017; Rudin, 2019; Samek et al., 2021).

We turn to deep neural networks (DNNs; Samek et al., 2021) to illustrate an example of the opacity of black-box models. DNNs are a family of algorithms that make decisions typically through a complex series of nodes. Nodes are combinations of input data with a set of weights or coefficients that either increase or decrease that input. There are often so many nodes in the algorithm that a human cannot understand all the information, even if it was provided, making them theoretically explainable but not necessarily interpretable or understandable (Angelov and Soares, 2020). Given their complexity, they do not provide meaningful information about the algorithm's underlying decision processes to the user (see also Samek et al., 2021). This resulting lack of human understanding as to the underlying decision-making complexities has led researchers (Barreto Arrieta et al., 2020; Sanneman and Shah, 2022; Vilone and Longo, 2021) and governments (European Union Act, 2024/1689; Air Force Doctrine Note, 2024) to call for transparency in these algorithms as they are leveraged to make key decisions.

There is a tradeoff between white- and black-box models. Black-box models are often the most predictive / accurate, yet they are the opaquest, whereas white-box models provide all or most of the information used in reaching a given classification but are less accurate (Herm et al., 2023). Differences between black- and white-box models are non-linear with numerous parameters that are not easily interpreted (Li et al., 2022; Mahmud et al., 2021). However, the tradeoffs for black- and white-box models are not as straightforward as previously thought, such that there is nuance between models (Herm et al., 2023). Traditionally, researchers and practitioners have needed to balance the tradeoff between the performance and explainability of these algorithms (Rudin, 2019). The advent of XAI in the last 20 years has attempted to alleviate some of the explainability issues in black-box models, applying these algorithms to highlight the important features of a black-box model's decision-making process to afford human understanding (Adadi and Berrada, 2018; Herm et al., 2023).

XAI

XAI are algorithms that make it possible for humans to keep intellectual oversight of AI/ML (Adadi and Berrada, 2018; Gunning and Aha, 2019; Longo et al., 2024). The focus of the literature surrounding XAI is to create algorithms that provide explanations for AI/ML decision processes in a manner that is interpretable for human users (Ali et al., 2023; Sanneman and Shah, 2022; Visser et al., 2023; Hassija et al., 2024). XAI explanations are meant to facilitate appropriate reliance and proper use of AI/ML systems, ensure fairness in the resulting decisions informed by those systems' outcomes, and provide an understanding of where these systems are lacking in performance (Barreto Arrieta et al., 2020; Visser et al., 2023). However, even the concepts of explainability versus interpretability within the XAI literature are nuanced, and the unique challenges for affording both are not necessarily one and the same (Guidotti et al., 2018; Tocchetti and Brambilla, 2022).

A key aspect to XAI is the theory that if users can interpret the behavior of the algorithm, whether correct or incorrect, they will be more willing to act on the suggestions of the algorithm

2 We note that several taxonomies in the computer science literature exist for demarcating the different types of algorithms, e.g., Barreto Arrieta et al. (2020), Speith (2022), and Zhou et al. (2021). However, the focus of the current work is facilitating proper cognitive evaluations of the systems. As Alarcon and Willis (2023) note, many of the taxonomies in the literature do not result in differences users can detect.

appropriately, especially in instances where predictions are not consistent with the user's expectations (Berger et al., 2021; Ribeiro et al., 2016b). In other words, explanations can bridge the information gap between the AI/ML model and the user (Baird and Maruping, 2021; Barreto Arrieta et al., 2020). However, research has found less than 1% of XAI studies in the literature contain user interactions with these models (Keane and Kenny, 2019; Suh et al., 2025).

Current paper

The research on human interpretations of AI/ML/XAI has been largely atheoretical in previous research. Black-box models were created largely without the user in mind as much of the early research was focused on model accuracy. The advent of XAI has sought to remedy these limitations but has not focused on human-machine interactions (Keane and Kenny, 2019; Suh et al., 2025). Although research has been conducted on increasing interpretability of AI/ML/XAI models, there remains a relative dearth of theoretical frameworks for creating these models. A comprehensive theory of how users comprehend AI/ML/XAI is necessary to help understand how and why users trust a system so that systems can be designed with the user in mind. The current paper bridges the gap between the human-computer interaction literature in social sciences and the AI/ML/XAI literature in the computer sciences. We sought to create a framework based on information processing theory in the social sciences and of AI/ML/XAI and experimentation to determine if the advances of new models meet the creator's criteria of facilitating trust. The theoretical model provided below provides key testable hypotheses for researchers when developing AI/ML/XAI for users and developers.

User-centered trust towards AI model

Deciding to trust or rely on a machine is inherently an information processing model (Alarcon et al., 2023; Chiou and Lee, 2023), where system transparency leads to more information about the system that should inform the user and thereby facilitate calibrated trust, assuming the appropriate amount of information—both its perceptibility and veridicality—is displayed.

Although we do view much of the trust process as previous researchers do, we expand on the previous research explicating the mechanisms for many variables in the model. Figure 1 illustrates our theoretical model of trust in human-AI/ML/XAI interaction. Information is perceived from the environment, processed by the user, and mental models of the system are formed. Information below the black line illustrates aspects of the machine, which have been drawn with squares. Aspects of the model above the black line illustrate aspects of the user, which are drawn with circles. Information processing occurs on the black line, with the user interpreting information about the system as illustrated with grey rounded boxes. The mental models about the referent lead to attitudes about the system which result in behaviors.

In our theoretical model, we demarcate between antecedents to trust, trust, and behaviors based on previous models (Hoff and Bashir, 2015; Lee and See, 2004; Mayer et al., 1995; Schlicker et al., 2025; see also Kohn et al., 2021). Antecedents to trust in our model comprise individual differences and trustworthiness perceptions, much like previous literature (Kohn et al., 2021).

Moving beyond previous research, we expand on the various individual differences and their theoretical effects on the trust process. We also expand on the process of transparency, describing how transparency of the referent machine model influences cognitive processing of information in the environment and links to relevant literature in the computer science research, as illustrated in Figure 2. Information from the environment is processed and comprises trustworthiness perceptions, which influence attitudes toward the system. Importantly, we do not theorize trust is the only antecedent to behaviors. We expand on the different attitudes that can influence behavioral outcomes with a system.

Figure 1 incorporates all the previous research that we have described into one cohesive model. First, we note that unlike other models (but see Schlicker et al., 2025), there is a clear demarcation between the user and the machine. The bottom of the figure illustrates the AI/ML/XAI algorithm and the information it provides, which we call trustworthiness cues. The upper half of the image illustrates the information perceived by the user, which we call trustworthiness perceptions. This was done to emphasize the difference between objective aspects of the machine referent and the user's perceptions of the machine. Although previous theoretical models implied the demarcation of trustworthiness cues from trustworthiness perceptions (e.g., Alarcon and Willis, 2023; Hoff and Bashir, 2015; Lee and See, 2004), we explicitly state and illustrate these differences in our model. We have used squares to represent objective information, circles to represent subjective perceptions, and squares with rounded edges to illustrate the utilization/processing of information in the figure (see Figure 1). Table 1 defines all constructs and provides citations for each construct.

Individual differences and information processing

The information processing approach to human perception and cognition has been utilized to much acclaim in the Human Factors literature (Wickens and Hollands, 2000; Wickens, 2002; Wickens and Carswell, 2012). However, the role of individual differences in the information processing theory has largely been ignored (Endsley, 2023; Wickens and Carswell, 2012). In this section we outline the types of individual differences, and the three proposed influences individual differences can have on information processing: (1) individual differences as information, (2) individual differences in the processing of information, and (3) individual differences as direct effects on behaviors.

There are individual differences between users, which can theoretically explain different perceptions of and interactions with machine systems in general (Hoff and Bashir, 2015; Lee and See, 2004) and AI/ML-enabled systems specifically (Kaplan et al., 2023). We demarcate these into general beliefs, personality variables, and demographics. We theorize individual differences play a role across all aspects of the decision-making process when humans interact with AI/ML/XAI. We expand on this throughout the theoretical model. In this section, we discuss the three types of individual differences; then, we discuss the three mechanisms through which they operate on the trust process.

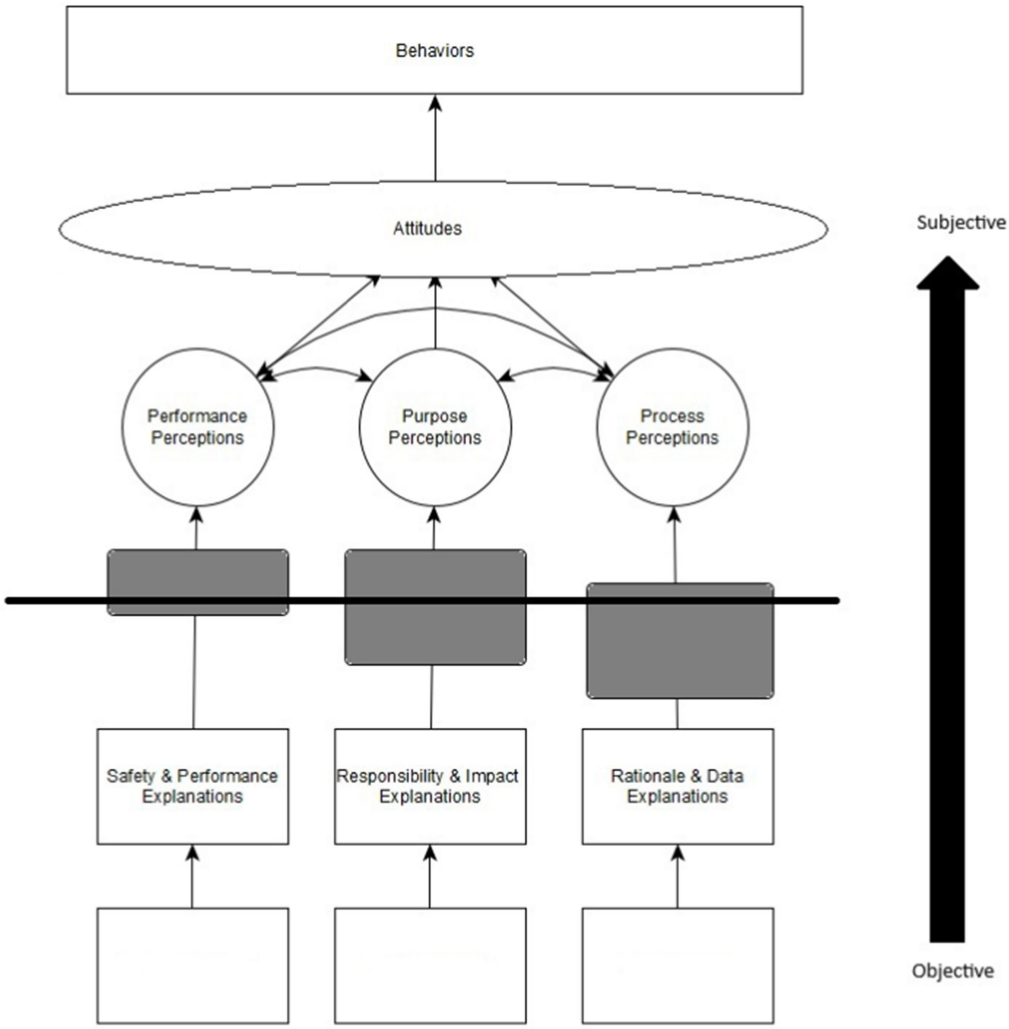

FIGURE 1
User-centered trust towards AI model. Objective aspects of the model are illustrated as squares. Subjective aspects of the model are illustrated as circles. Subjective user interpretation of objective information is illustrated as rounded boxes. The figure illustrates the information from AI/ML models in the form of trustworthiness cues. Trustworthiness cues are then interpreted by the human, and the demarcation line through the rounded boxes denotes the emphasis on the AI/ML referent (below) or user (above). The subjective interpretation of the referent system leads to subjective trustworthiness perceptions. These perceptions facilitate attitudes toward the AI/ML-enabled system. The attitudes facilitate behaviors. Trustworthiness cues were left blank as they depend on the models used.

Global beliefs

Global beliefs can influence perceptions through a variety of variables. We define global beliefs as any general belief or cognitive schema about AI/ML/XAI that are preconceived *prior to working with* a specific AI/ML/XAI referent. Perhaps the two most researched global beliefs variables in relation to the human-machine trust process are propensity to trust machines (Jessup et al., 2019; Li et al., 2017; Merritt, 2011) and perfect automation schema (Dzindolet et al., 2002; Gibson et al., 2023; Merritt et al., 2015). Propensity to trust machines is defined as a general tendency to utilize machines or view them favorably (Jessup et al., 2019). The propensity to trust machines is a global heuristic about machines in general that influences initial perceptions of a new system particularly when there is little information about the system available to the user (Hoff and Bashir, 2015; Siau and Wang, 2018). Perfect automation schema is like propensity to trust machines in that it pertains to global perceptions of machines; however, this construct comprises high expectations and all-or-none thinking (Merritt et al., 2015). High expectations are the initial beliefs that machine systems should perform well and thus has strong overlap with propensity to trust machines (Gibson et al., 2023; Merritt, 2011). All-or-none thinking is the bias that machine systems should be abandoned if they commit an error.

Importantly, these are all global beliefs about systems which are informed by experience with the world, rather than personality constructs. We note these two global beliefs are not theorized to be an exhaustive list of beliefs but rather just two examples of beliefs in the literature.

Personality

Personality is defined as any cognitive, emotional, or behavioral patterns that comprise an individual's unique perspective in life.

TABLE 1 Construct definitions and example reference sources.

Construct	Definition
Individual differences	
Global beliefs	Beliefs, attitudes, or schemas about machine systems *in general* (e.g., Hoff and Bashir, 2015)
Personality	Cognitive, emotional, or behavioral patterns that compose a user's unique perspective (e.g., Matthews, 2008)
Demographics	Population information (e.g., age, occupation) about the user (e.g., Hoff and Bashir, 2015)
Transparency	
Trustworthiness cues	Objective output from the referent machine system (e.g., Zhou et al., 2021)
Trustworthiness perceptions	A user's subjective perceptions of the referent machine system (e.g., Schlicker et al., 2025)
Performance	Perception of a machine system's capabilities (e.g., Lee and See, 2004)
Process	Perception of the algorithmic steps in which a machine system operates (e.g., Lee and See, 2004)
Purpose	Perception of whether a machine system is being used as its designer intended (e.g., Lee and See, 2004)
Explainability	Objective information provided by a machine system (e.g., model weights, estimates) represented through visual, textual, or numerical data (e.g., Zhou et al., 2021)
Interpretability	Human making sense of the visual, textual, or numerical information provided by the machine system (e.g., Sanneman and Shah, 2022)
Understanding	Human cognitively processing the information from their interpretation to form schemas or perceptions of the machine referent (e.g., Sanneman and Shah, 2022)
Attitudes	
Trust	A willingness to be vulnerable (e.g., Lee and See, 2004; Mayer et al., 1995); a positive affective state coupled with appropriate effort
Distrust	An *un*willingness to be vulnerable (e.g., Lee and See, 2004; Mayer et al., 1995); a negative affective state coupled with appropriate effort
Suspicion	A state characterized by uncertainty, high cognitive activity, and perceived mal-intent (e.g., Bobko et al., 2014); a negative affective state coupled with high cognitive effort
Curiosity	A temporary state to seek more information (e.g., Kashdan and Roberts, 2004); a positive affective state coupled with high cognitive effort
Complacency	Decreased attentiveness resulting in relying on a machine system *beyond* its objective capabilities (e.g., Parasuraman and Manzey, 2010); a positive affective state coupled with low cognitive effort
Under-reliance	Relying on a machine system *less than* its objective capabilities can support (Lee and See, 2004); a negative affective state coupled with low cognitive effort
Risk	Perceived contextual uncertainty surrounding an outcome, and the relative vulnerability assumed surrounding trusting a system (e.g., Kohn et al., 2021).
Reliance/compliance behavior	Observed behaviors; abstaining behavior absent a machine system's suggestion and behaving when cued, respectively (e.g., Meyer et al., 2014)
Errors	Incorrect system processes or outputs
Unintended consequences	Results, either positive or negative, from a machine system not originally expected

Personality psychology is focused on individual differences in mental processes and how they develop (Roberts and Yoon, 2022). There are several personality structures in the literature that can influence human perceptions and behaviors, such as the Five-Factor Model (Costa Jr and McCrae, 1992). We note that relatively little research has been conducted on the effects of personality on human-AI/ML/XAI interaction (Kaplan et al., 2023). However, personality has been related to aspects of the information processing model. For example, the Five-Factor Model has been related to aspects of information processing theory such as working memory (Waris et al., 2018) and short-term memory (Matthews, 2008). Indeed, research has noted that task and person characteristics should not be explored in isolation but instead explored concurrently to determine their effects (Matthews, 2008; Szalma, 2008, 2009; Szalma and Taylor, 2011). As such, personality variables may help explain individual differences in how users process information from the system.

Demographics

Recent meta-analytic work shows the importance of demographic variables on human trust toward AI/ML (Ehsan et al., 2024; Kaplan et al., 2023).[3] Knowledge, skills, and abilities (KSAs) of the human user play a key role in understanding how they perceive AI/ML-enabled systems (Hoff and Bashir, 2015; Schaefer et al., 2016). For example, Ehsan et al. (2024) found novices had different heuristics for XAI than computer science majors, leading to different interpretation of the stimuli provided by the XAI. The age of the user may also be important in cognitive processing. Research has noted that younger users are more likely to have a positive view of technology and are more likely to use technology than

3 We note the authors of the meta-analysis did not explore XAI, which is why we did not use the term AI/ML/XAI.

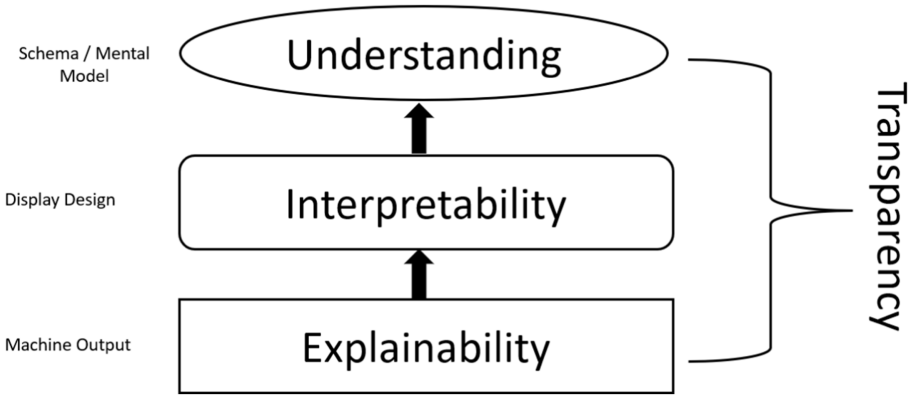

FIGURE 2
Transparency process as a function of explainability, interpretability, and understanding. We match the terms explainability, interpretability and understanding from the computer science literature to relevant constructs from the social sciences. Explainability is the machine output of relevant criteria (i.e., weights, estimates, etc.) in terms of textual, visual or numerical representation from the model. Interpretability is the cognitive processing of information by the user of the machine output. Understanding is the schema/mental model that is formed about the referent machine that facilitates attitudes in the system.

older individuals. This effect has been noted in a variety of research. For example, Casey and Vogel (2019) found millennials were more likely to use technology than any other generation. Males have also demonstrated a stronger propensity to use technology than females across a variety of situations. Thus, we see the import of simple demographic variables normally gathered in research as playing a role in baseline interactions with machine human systems (see Kaplan et al., 2023).

Mechanism for individual differences on perceptions

Individual differences have demonstrated the ability to act as information when little to no information about the referent or situation is present. Specifically, global beliefs such as propensity to trust have been theorized to act as information when little is known about the referent partner in interpersonal research (Alarcon et al., 2016, 2018; Alarcon and Jessup, 2023; Jones and Shah, 2016). A person that has a global belief that people are trustworthy will be more trusting of others when information is lacking. As information becomes salient about the referent, the loci of information transitions from the trustor to the trustee (Alarcon et al., 2016; Jones and Shah, 2016). We theorize similar relationships in human-AI/ML/XAI interactions.

Specifically, global beliefs can act as information about the environment or referent AI/ML/XAI algorithm when there is little to no information available. Demographic variables can also play a role in this initial perception. For example, Alarcon et al. (2017) found computer programmers were reticent to trust code from an unknown source, specifically if the use case was high in vulnerability. The training, experience, and personality variables the user has can act as information about the situation when information about the referent is lacking. The differences they found may be due to different cognitive heuristics developed through training and expertise (Ehsan et al., 2024). This information that is inherent in the user facilitates initial perceptions about the referent AI/ML/XAI.

> *Proposition 1: Global beliefs about AI/ML/XAI, demographic variables, and personality influence initial perceptions of the referent system.*

> *Proposition 2: As information about the referent AI/ML/XAI-enabled system is made salient over time, the influence of global beliefs will have less of an impact in the human-AI/ML/XAI interaction.*

Second, individual differences can influence the processing of information as the user receives information from the AI/ML/XAI output. As information becomes salient in the interaction with the AI/ML/XAI, the user will process this information which can be influenced by individual differences. First, global beliefs and personality differences can influence the processing of information. These constructs may influence the cognitive processing of information, and include personality variables that influence one's perceptions such as need for cognition (i.e., the extent to which individuals enjoy and engage in effortful cognitive activity; Cacioppo et al., 1996; Cacioppo and Petty, 1982), curiosity (i.e., the general desire for knowledge, to resolve knowledge gaps, to solve problems, and a motivation to learn new ideas and engage in effortful cognitive activity; Berlyne, 1960, 1978; Loewenstein, 1994), and complacency potential (i.e., a general propensity to *not* engage in effortful thinking; Merritt et al., 2019; Singh et al., 1993), along with global beliefs like all-or-none thinking. Other personality characteristics such as neuroticism, extraversion, and risk aversion can also play a role. For example, people high in neuroticism tend to view the world negatively (Bunghez et al., 2024), which can influence their perceptions about the AI/ML/XAI, especially if stimuli are ambiguous. Conversely, people high in extraversion tend to experience high positive affect (Costa Jr and McCrae, 1992), which may influence their perceptions of AI/ML/XAI. Alarcon and Jessup (2023) found risk aversion, a general disinclination towards risk, influences the processing of information in a modified version of the trust game. Participants high in risk aversion were less apt to view their referent partner as more trustworthy and took fewer risks with their partner.

Importantly, demographic variables such as knowledge, skills and abilities can also influence the processing of this information (Hoff and Bashir, 2015). For example, expertise can influence how information from the model is perceived through previously formed heuristics (Ehsan et al., 2024). Take the extant work

investigating algorithm aversion which is a reluctance toward relying on algorithms for decision-making compared to humans, even if the algorithm is more accurate than a human assistant (Capogrosso et al., 2025; Dietvorst et al., 2015). The research paradigm by Dietvorst et al. (2015) on algorithm aversion has noted that people are reticent to use algorithms in a GPA prediction task. Jessup et al. (2024) have noted algorithm aversion was strong for individuals who engaged in tasks with the GPA prediction as the focus, but if the focus of the task was classifying Github repositories, participants over-trusted the algorithm. Jessup et al. (2024) also note that participants may not have had the requisite knowledge, skills, or abilities in the Github repository task. Ribeiro et al. (2016b) note that interpretability of algorithm is dependent on the target user. Thus, the reason for algorithm aversion's effect in human-machine interaction may be due to the individual's knowledge of, skills pertaining to, and abilities using AI/ML-enabled algorithm systems.

> *Proposition 3: Users' global beliefs, personality, and demographics will influence their processing of information, thereby impacting their perceptions of and attitudes toward AI/ML/XAI referents.*

Lastly, we theorize that some individual differences can have a direct impact on behaviors. Personality variables such as complacency potential, curiosity, and need for cognition may all have direct influences on behaviors because of the nature of the constructs. As noted in the Psychology literature, behaviors are the outcome of many different psychological processes. In this instance, whether a user does or does not perform monitoring behaviors may be indicative of personality rather than aspects of the system (Gibson et al., 2023). Notably, need for cognition is typically associated with positive emotions under the broaden and build theory of emotions (Fredrickson, 2001), which can further lead to greater exploration. This exploration can lead to increasing or decreasing trust behaviors in human-AI/ML/XAI interactions, depending on the results of the exploration. There may be contexts in which combinations of these variables amplify, suppress, or have no effect on trust-relevant criteria of interest, which we discuss later in the Attitudes section. For example, someone with a high complacency potential may be more likely to rely on an AI/ML system at baseline and more so given contextual factors (e.g., time pressure, system opacity; Singh et al., 2021). However, if one has a high need for cognition and enjoys thinking deeply, they may be more apt to spend time trying to figure out the underlying processes by which an AI/ML system serves their needs in said context, regardless of how the system operates. This is just one example of these variables interacting to shape human-AI/ML interaction, and their effects likely have different impacts given the context.

> *Proposition 4: Individual difference variables associated with high or low cognitive engagement will directly influence a user's engagement with AI/ML/XAI referents.*

Transparency

There remains confusion in the XAI literature between explainability, interpretability, and understanding often with the terms being used interchangeably (Guidotti et al., 2018; Tocchetti and Brambilla, 2022). We view all of these as the process of establishing transparency. Transparency as typically referred to in the literature is related to the user's ability to detect how a referent system operates based on cues of the system (Lyons, 2013). However, transparency is also treated as a subjective construct in the literature (Alarcon and Willis, 2023; Chiou and Lee, 2023). Transparency pertains to the degree to which human-users are *able to perceive* system performance (what is the system doing), purpose (is the system being used as it was designed to be used), and process cues (how is the system doing what it is doing) as information (Chiou and Lee, 2023). A system that is transparent to one user may be opaque to another user because they lack specific knowledge, skills, or abilities (Ehsan et al., 2024; Jessup et al., 2024). Below we demarcate explainability, interpretability, and understanding within the context of transparency, with explainability being the most objective form of transparency, interpretability being a mix of objective information and subjective cognitive processing, and understanding being mainly a subjective perception or schema, as illustrated in Figure 2. In other words, transparency is the process of establishing ascriptions of an AI/ML-XAI-enabled system through information provided by cues in the environment.

Explainability—machine transparency and trustworthiness cues

An explanation is something that is provided by a referent, for instance, the details of an AI/ML-enabled system by an XAI tool or beta weights provided by a regression. Explainability is a representation of aspects of the algorithm, such as weights or noting important aspects of the stimuli, through visual, textual, or numerical representation. Explanations are a set of details or elements that illustrate or elucidate the causes, contexts, or outcomes of those details or elements (Drake, 2018). In our model, the trustworthiness cues provide different types of explanations. Importantly, the information provided by the trustworthiness cues may provide more than one type of explanation.

As Alarcon and Willis (2023) note, most of the taxonomies in the Computer Science literature refer to the underlying mathematical processes for computation. The different AI/ML/XAI taxonomies all attempt to classify the different algorithms utilized to create the various AI/ML/XAI methods. We acknowledge these differences are important as they help to describe how the various AI/ML (e.g., decision trees, DNNs) and XAI (e.g., local interpretable model-agnostic explanation, or LIME) algorithms are computing different weights and calculations for their respective predictions, classifications, or descriptions of a given data set. However, the individual model taxonomies are all classifications of how the algorithms function rather than the information provided to the user. The algorithms that are being created are objective, in that they provide some information about their inner workings (or no information in the case of black-box models) and then display this information to the user. The method used for analyzing what was important in the algorithm may vary, but often they result in similar output such as numerical, visual, or text output (Zhou et al., 2021). The differences of the model methods are not necessary for most users. It is the meaning of the numerical, visual, or text output within the context that is important for interpretations (Broniatowski and Broniatowski, 2021). As such, we placed the different AI/ML/XAI taxonomies for model methods in the squares at the bottom of the model in Figure 1, as they are the most objective.

The taxonomy provided by Zhou et al. (2021) differentiates algorithms by their explanation types as described in the previous section. Although they intended their taxonomy to only apply to XAI methods, we theorize it can be extended to all AI/ML/XAI methods *and* algorithms. The explanation types provide information to the user which is displayed in various manner. Black- and white-box models both adhere to these classifications, as the former provides little information besides performance information and the latter are fully or almost fully transparent as to their processes. These explanation types are similar to previous demarcations of objective referent trust from subjective trust perceptions (Schlicker et al., 2025).

Zhou et al. (2021) demarcated six explanation types: rationale, data, responsibility, impact, fairness, and safety-performance. Their study focused on the type of explanation the algorithm was providing to the user. The rationale explanation focuses on explaining why the algorithm made a specific decision to the user. For example, a LIME algorithm output can provide rationale for a certain classification by highlighting the relevant information. Data explanations illustrate what contextual data the algorithm used to make its decision. Figure 3 illustrates an image that has been classified as a civilian airplane, and the LIME data explanation highlights the relevant information for the classification in green.

Responsibility explanations focus on the development of the algorithm, how the algorithm was managed, and how the algorithm was implemented. This type of explanation is focused on the accountability of the algorithm and who has ultimate responsibility for the implementation of the algorithm. Responsibility explanations have largely driven laws regarding AI/ML in Europe (Hoofnagle et al., 2019). Impact explanations illustrate the societal impact of utilizing algorithms and possible consequences of using them in certain arenas. For example, researchers and governments have discussed the possible implications of self-driving cars (which utilize AI/ML for their autonomous function; Holstein et al., 2018). Self-driving cars will have various impacts on the legal system, ethical decision-making, and loci of responsibility (human or machine), to name a few.

Fairness explanations can be viewed as a subset of impact explanations as they are specifically focused on possible bias pertaining to their use. As mentioned, AI/ML has been utilized in parole decisions in recent years. However, these algorithms have been shown to be biased against minorities based on various variables such as zip codes (Hübner, 2021). The impacts of these biases and the broader implications of using AI/ML in sensitive areas have been discussed at length by both researchers and governments (Goodman and Flaxman, 2017; Hoofnagle et al., 2019). Lastly, safety and

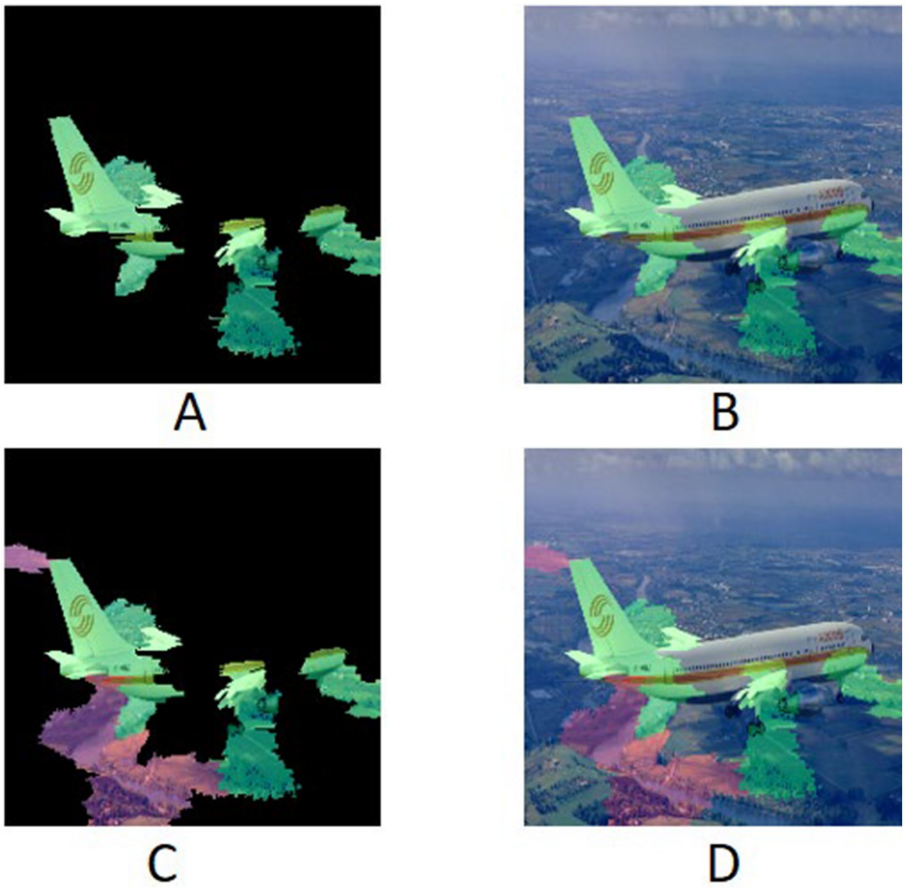

FIGURE 3
Locally interpretable model-agnostic explanation of civilian airplane. **(A)** Obfuscated image with positive information highlighted; **(B)** non-obfuscated image with positive information highlighted; **(C)** obfuscated image with positive and negative information highlighted; **(D)** non-obfuscated image with positive and negative information highlighted.

performance explanations illustrate the process of increasing the accuracy, security, reliability, and robustness of the decisions of and output from AI/ML algorithms. An example of this is when a researcher finds there are data in the data set that the algorithm has not been trained on or that the use of said data has unintended consequences. For example, Target utilized AI/ML for targeting coupons toward their customers. In one instance, the algorithm noticed a 16-year-old girl was pregnant and sent her various coupons for baby supplies, angering her father who was unaware of her pregnancy. As such, Target recalibrated the AI/ML to only offer coupons to those over 18 years of age (Oprescu et al., 2020).

The explanation types described by Zhou et al. (2021) are the outputs of the AI/ML/XAI algorithms, which we illustrate in Figure 2. We note that even a black-box model will have trustworthiness cues related to it, but in comparison to an XAI or white-box model, there will be relatively fewer as the traditional black-box models do not typically illustrate much information other than performance in their output (Adadi and Bouhoute, 2023; Vilone and Longo, 2021). In other words, black-box models are simply low in process cues as they do not illustrate their decision processes for humans. Alarcon et al. (2024) note that all algorithms lie somewhere on the information spectrum and that a lack of information, such as with black-box DNNs, does not negate the theory.

The explanation types are closest to the objective reality of the algorithm in our theoretical model. AI/ML/XAI often comprise weights and errors for the algorithm but can be illustrated differently through natural language, color, or text (Zhou et al., 2021). It is how the information is displayed that facilitates effective human processing of the information. To illustrate this point, we use another example of LIME algorithms for image classification. There are several different ways to view the output of a LIME algorithm. Utilizing graphical images, we can increase the interpretability of the XAI output. Figure 3 illustrates several LIME outputs from the same algorithmic output. First, Figures 3A,C present LIME output that illustrates the important aspects of the image in making its classification of the stimuli on a black background. In this instance, LIME XAI displays the relevant information, but the context is not clear. In Figures 3B,D, the LIME XAI is transposed on top of the image, with translucent color so the relevant aspects of the image and the total image are clear. In this example, the user may be able to better understand the aspects of the image that helped the AI/ML classify the image. Third, Figures 3C,D offer two types of relevance to the image classification, green which is associated with information that influences the classification as civilian, and red which illustrates information that is not relevant to the classification. All these outputs were based on the same algorithmic output but are displayed differently.

Zhou et al. (2021) do not explicitly state whether the explanation types are orthogonal or not. We, however, contend these are *not* orthogonal. The example of the Target shopper above illustrates this point: the AI/ML sending the coupons to an underage person contributes to both impact and performance explanations. The performance explanation was correct, in that the daughter was pregnant. However, the impact explanation of sending the coupons to anyone under the age of 18 was also clear from the output of the algorithm. Thus, the same output can have different impacts based on different explanation types. We note that there is no influence of individual differences on the trustworthiness cues as they are inherent in machines. The output is the objective information but *how* that information is displayed is important from the human perspective, as noted in the LIME XAI examples above. We do acknowledge that user inputs at the algorithm level can occur, but these are more important when creating and training the algorithm.

> *Proposition 5: AI/ML/XAI-enabled systems that provide more information on their underlying processes pertaining to their respective explanation types will be viewed as more informative.*

Interpretability—cognitive processing of information

We define interpretability as a human making sense of the information provided by the AI/ML-enabled algorithm or XAI, as illustrated in Figure 2. In the human-AI/ML/XAI interaction context, it is the human that ascribes meaning to the information provided by the algorithm. Just because information is explained does not mean it is interpreted or understood by the user. How the information is displayed helps the user interpret the data. As such, per Audi (1999) we define interpretation as the mental processing of information provided from the system (i.e., explained information) by the user to establish an (in)accurate understanding of the system referent (e.g., Ribeiro et al., 2016b). Indeed, Ehsan et al. (2024) found differences in interpretation of XAI between novices and computer science students with the same XAI stimuli.

To illustrate our point, Figure 3 illustrates a LIME algorithm representation of the information that led to the classification of a civilian airplane with the relevant weights being highlighted in green and red. Green illustrates the top five most relevant aspects of the image to classify it as a civilian plane. Red illustrates the five least relevant aspects of the image to the classification as a civilian plane. Although these are illustrated as colors over the image, it is the numerical value that the algorithm weights for each pixel that is important. The LIME algorithm describes the visual information that is used to generate a classification of each aspect of the image. The data being explained by the machine are pixels. The color is a quick visual cue for easy information processing for humans to interpret (Xing, 2006). As such, if the algorithm simply listed the pixels used in making the determination, the information would be uninterpretable but explainable because there is too much information for users to process (e.g., Young et al., 2015).

In Figure 1, the line across the middle of the figure illustrates the demarcation of subjective user perceptions and objective machine output. In the lower portion of the figure, performance, purpose, and process cues from the machine are illustrated with square boxes. There is a direct path from the trustworthiness cues to the user's trustworthiness perceptions, which are noted in the model with performance, purpose, and process perceptions, instantiated with circles to indicate subjective perceptions. Although the current figure only illustrates direct lines from each trustworthiness cue to its respective trustworthiness perception, this is only done for clarity of the theoretical model.

The grey boxes on the lines between the trustworthiness cues and perceptions are the information the user is employing to interpret the trustworthiness cues (Schlicker et al., 2025). Individual differences are an integral part of how the user perceives and interprets the information. For example, Ehsan et al. (2024) found participants in a computer science program perceived information provided from an AI/ML algorithm differently than novices. In the first line from performance cues to

performance perceptions, the grey box is mostly on the upper portion of the figure indicating the user is relying on information within oneself (e.g., personal beliefs, experience with AI/ML) to make the assessment rather than the cues from the machine. This illustrates an instance when individual differences such as propensity to trust technology or previous experience with AI/ML will be driving the decisions, as the loci of information for the assessment resides primarily in the user and not the system. This is what Hoff and Bashir (2015) refer to as dispositional trust. The grey box representing the performance cue information the user is perceiving is smaller than the boxes for purpose and process, representing less trustworthiness information when making the decision because they are relying on individual beliefs.

> *Proposition 6: Users deferring solely to their global beliefs / cognitive schemas to inform their trustworthiness perceptions of AI/ML/XAI will understand less about these systems.*

On the line from the purpose cues to purpose perceptions, the grey box indicating information for perceiving the system is split hallway between the user's perceptions and the information from the algorithm. In this instance, we also see the box is much larger as more information is being utilized, with equal information from the user and the machine. The grey box on the purpose line might indicate an expert user with adequate domain knowledge is properly utilizing balanced information from both subjective perceptions and objective cues in the environment. This example would be representative of a statistician that understands and is utilizing regression weights in making an informed decision in a domain for which these models were designed to be applied.

> *Proposition 7: Users who leverage their domain knowledge coupled with relevant information from AI/ML/XAI systems will more appropriately calibrate their trust toward the system.*

Lastly, process perceptions are illustrated with a grey box with most of the variance for information being held on the objective side of the figure. In this instance, the algorithm may be providing information that requires little to no interpretation by the user. We can think of image classification for easy images which are often used to develop new models, such as classifying animals, as they require little cognitive processing from the developer. Across Figures 3A–D, we can see that the algorithm primarily utilizes information about the airplane in its classification. However, there are aspects of the image background that also are included in making the determination. Figures 3A–D would be representative of the box on the process line because it requires little cognitive processing by the user, as most users would be familiar with and comfortable classifying the image as a plane. These three examples illustrate the differences in interpretability across individual differences.

> *Proposition 8: Users leveraging objective information solely from the AI/ML/XAI will benefit in their interaction with the AI/ML/XAI system so long as that information provided is veridical.*

Some important aspects of the performance, purpose, and process perceptions should be discussed. First, XAI may not reduce overreliance on AI/ML (Müller et al., 2024). Instead, it may *increase* overreliance as research has demonstrated users are more likely to agree with an AI/ML algorithm if it provides an explanation, regardless of accuracy (Poursabzi-Sangdeh et al., 2021). Indeed, Ehsan et al. (2024) found novices trusted the numeric output of an algorithm simply because it was numeric, inferring that the numbers were based on algorithmic thinking. However, this may be moderated by personality variables such that those with greater attention to detail may not over-trust. Situational variables such as workload can also influence trust as users that have too many demands placed on them may over-trust because of lack of monitoring resources (Biros et al., 2004). The information provided by the algorithm can facilitate many aspects of understanding. Colin et al. (2022) found attention maps helped facilitate user understanding of biases in the AI/ML algorithm, but it did not facilitate understanding of the failure cases. As such, information provided by the algorithms may not facilitate a full understanding but rather understanding of different outcomes.

Understanding

The user exploits the textual, numeric, image, or natural language output to facilitate their understanding of the decision process of the algorithm (Roscher et al., 2020). It is the *ability of the human to cognitively process* the output of the algorithm that leads to understanding (i.e., cognitive perceptions). Thus, understanding is the perception of the system as a result of the interpretation of the explained information. If the algorithm does not display the information in a format that is interpretable to the user, or if the user does not have the requisite knowledge, skills, and abilities to interpret, it may have high explainability, low interpretability (i.e., the user is not able to make sense of the explained output), and lead to misunderstanding (i.e., a misperception of what the model is explaining, how it functions etc.). However, if the user does have the requisite knowledge, skills, and ability, this can facilitate appropriate sense-making of the data and ultimately properly calibrated understanding of the system (Ehsan et al., 2024; Klein et al., 2006). This expression of the information facilitates interpretation which enables an understanding about what will happen in the future, as illustrated in Figure 2 (Koehler, 1991; Lombrozo and Carey, 2006; Mitchell et al., 1989). These perceptions result in an understanding of the stimuli, which is highly subjective. Thus, understanding is the schema or mental model that is created by the user from interpreting information in the environment, which can be used as a lens of analysis in future interactions with AI/ML and XAI systems. Typically, these schemas and mental models are assessed by measures of users' performance, purpose and process perceptions of machine systems (Lee and See, 2004; Stevens and Stetson, 2023), in this case AI/ML/XAI.

> *Proposition 9: Knowledge, skills, and abilities (KSAs) will be important for facilitating understanding, such that information from the AI/ML/XAI should be displayed in relation to the users' KSAs.*

As noted above, the construct of understanding from the computer science literature is synonymous with trustworthiness perceptions (comprised of performance, process and purpose) from the social sciences literature. *Performance* perceptions concern the degree to which the user perceives the system can perform a specific task within a given context (Lee and See, 2004), what Hoff and Bashir (2015) would term situational trust. Some machine systems may

be perceived to have good performance for certain tasks but not others. It may also be the context of the system's use that moderates situational trust. For instance, an AI/ML-enabled system may perform well at classifying images from classes it was trained but not out of distribution classes to perceive algorithm performance differently depending on how the algorithms are applied[4]. The risk of misclassification may be tolerable in scenarios such as recycling management but not in high risk instances such as computer vision for autonomous vehicles. *Process* perceptions describe the user's understanding of a machine system's underlying algorithmic function (Lee and See, 2004). In human-AI/ML contexts, process perceptions may vary depending on what information a user perceives on how the AI/ML reached a decision, and these perceptions may be shaped by XAI which unpacks how the AI/ML functions (Alarcon and Willis, 2023). *Purpose* perceptions pertain to a user's perceptions of why the system was designed (Lee and See, 2004). Lee and See note systems are often used outside of contexts they were built or used in contexts that have more diverse information than that which the model was trained on, which can influence users' purpose perceptions. With regards to the above example, systems designed to classify specific classes such as cats and dogs, but not other classes (e.g., squirrels), may shape users' perceptions of system purpose, which may or may not relate to other trustworthiness perceptions and intentions to rely on (i.e., trust) the system as shown in human-autonomy interaction work (Capiola et al., 2023; Lyons et al., 2021, 2023).

Providing information is not a catch-all that once provided will increase trust and reliance on the system. The explanations provided can reveal problems in the system which can help the user calibrate when to use the algorithm (Kästner et al., 2021). For example, Alarcon et al. (See footnote 4) found participants were able to more quickly detect out-of-distribution data and noted the open set recognition models were better able to classify appropriate images for within-distribution data; in comparison, the convolutional neural networks (CNNs) had issues because they were confident in their classifications regardless of the stimuli. Additionally, the user can misinterpret the information provided by the model, falsely attributing actionability (i.e., what can be done with the information) even when the explanation is unclear (Ehsan et al., 2024).

The explanation of the algorithm's decision processes when the algorithm fails can also provide information as to improvements that need to be made to the algorithm. For example, adversarial attacks are used to determine ways to deceive the algorithm. Researchers noted that adding black and white bars to the stop sign, as illustrated in Figure 4, can make a convolutional neural network classify the image as a 35-mph sign instead of a stop sign. This helps the developer find issues in the algorithm that can be alleviated or fixed with updates and understand the flaws in the algorithm and improve it in future iterations.

Attitudes

There are numerous definitions of trust in the AI/ML/XAI literature. Many of these definitions have come from the interpersonal trust literature and have also conflated trust and reliance. Blanco (2025) recently theorized trust comprises (a) positive expectations, (b) some risk of the trustee not behaving as the trustor wants them to (c) a potential or need / wish of delegation, and (d) a basis in motives (i.e., motives based). She distinguishes this from reliance, which is the dependence of the trustor toward the trustee when the trustor needs to delegate. In her definition, the first two postulates of trust are the same as most definitions in the literature. Her third postulate of trust distinguishes that needing to delegate is not a necessary condition of trust, but rather the potential wish for delegation can necessitate trust. Additionally, she explicitly notes that trust is motives based, i.e., trustors have a goal in mind when ascribing trust. In her theoretical paper, she notes these motives are based on both cognitive (normative) and affective antecedents. Indeed, in the interpersonal trust literature,

4 Alarcon, G. M., Jessup, S. A., Meyers, S. K., Willis, S., Johnson, D., Noblick, J, et al. (under review). The trust process applied to machine learning algorithms: the influence of calibrated confidence estimates. Manuscript under review.

FIGURE 4
Example of adversarial attack on road sign classification task.

there has also been a demarcation of affect- and cognition-based trust (McAllister, 1995). However, the interpersonal trust literature has focused on differentiating antecedents to trust as affect- and cognition-based aspects of trustworthiness perceptions (Colquitt et al., 2007; Lee et al., 2022).

We theorize two underlying factors that influence attitudes toward AI/ML-enabled systems: valence and cognitive effort. Valence describes the attractiveness or aversiveness the user ascribes to the situation or referent (Russell, 1980). Valence comprises a spectrum of emotions ranging from positive to negative, with neutrality or indifference in the middle (Jessup et al., 2025; Li et al., 2017). Research on emotions has demonstrated that when users experience positive emotions, they tend to process information more holistically (Fredrickson, 2001; Kazén et al., 2015). In the human-machine interaction literature, this can facilitate adoption and acceptance of the system (Hoong et al., 2017; Jessup et al., 2023). Positive emotions also lead the user to explore the machine's capabilities, experiment with features, and engage more with the system (Isen and Geva, 1987). Additionally, positive emotions can buffer against system failures or errors (Maroto-Gómez et al., 2023). In contrast, negative emotions can facilitate narrow or segmented cognitive processing (Kazén et al., 2015). Initial negative interactions are difficult to overcome as they lead to a negative mental model of that referent system (Jessup et al., 2020), which takes time and effort to change (Kim et al., 2023).

Cognitive effort characterizes the mental exertion necessary to evaluate a referent. Though there is no consensus on a definition (Li et al., 2017; Shepherd, 2022; Westbrook and Braver, 2015), cognitive effort is often described as the amount of controlled processing needed to understand, reason, and make decisions. This overlaps with dual-processing models of cognition (Kahneman, 2011) and persuasion (Chaiken, 1980; Petty and Cacioppo, 1986), which have also been applied to trust research (Alarcon and Ryan, 2018; Stoltz and Lizardo, 2018). System 1 processing, or heuristic processing, is quick, automatic, and relies on cognitive shortcuts. System 2 processing is more effortful, intentional, and accurate. Importantly people bring heuristics about referents into the decision-making process based on knowledge domain, experience, and pattern recognition (Ehsan et al., 2024). Aspects of the environment such as task complexity, familiarity, and contextual distractions can also influence the amount of cognitive processing an individual utilizes (Wickens and Carswell, 2012). For example, research has demonstrated that people choose to exert cognitive effort when machine systems are coupled with explanations for their decisions by weighing the cost of the cognitive effort compared to simply deferring to the overall system (Vasconcelos et al., 2023). This is what Hoff and Bashir (2015) term situational trust. In addition, a user's motivation and cognitive ability (e.g., working memory capacity, processing speed) can influence the amount of cognitive effort a user is willing to exert on the task (Cacioppo et al., 1996). That is, cognitive effort facilitates or inhibits information processing, leading to resulting attitudes.

We theorize multiple attitudes toward AI/ML/XAI fall on a two-dimensional plane of valence and cognitive effort as they are a function of both dimensions, as illustrated in Figure 5. Lee and See (2004) are most famous for defining and explicating what it means to have *calibrated* trust, which is the correspondence between objective system capability and the user's subjective trust toward that system informed by their perceptions of that system. Trust calibration was originally introduced in the Human Factors literature concerning trust in automation and has since been adopted and expanded in the human-machine interaction literatures, including human-robot interaction (e.g., Alarcon et al., 2023), human-autonomy interaction (e.g., Capiola et al., 2023), and human-AI/ML interaction (e.g., Harris et al., 2024; Schlicker et al., 2025; for a review, see Kohn et al., 2021). In the present model, calibrated attitudes fall on the x-axis, as they

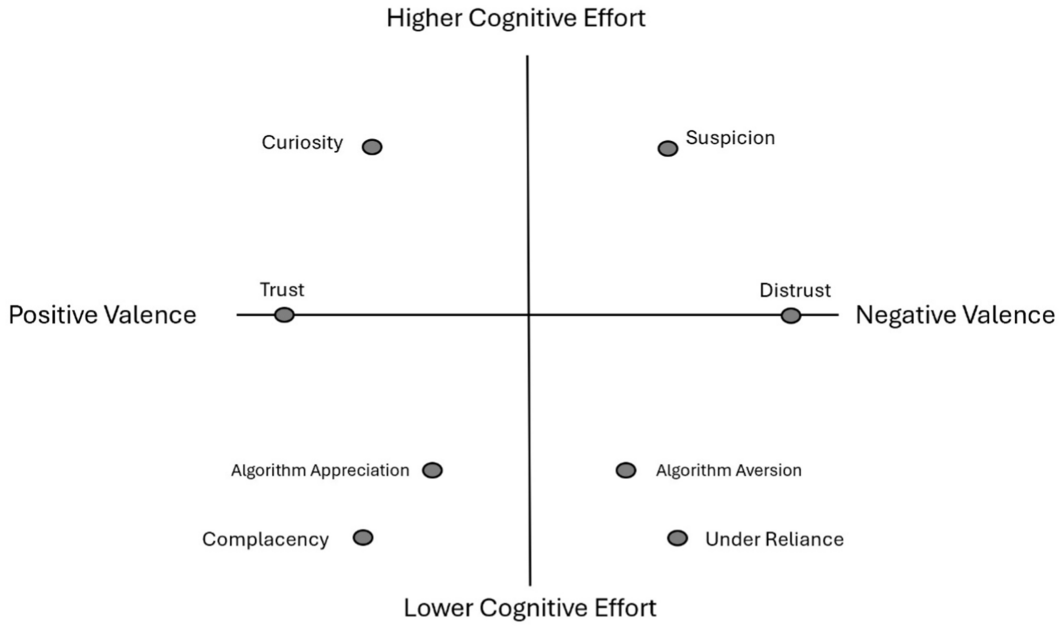

FIGURE 5
Trust and trust related attitudes along valence and cognitive effort factors. Contextual risk negatively affects attitudes; high and low workload contexts can affect processing resources and effort.

entail appropriate cognitive effort, and depending on the situation, comprise positive or negative valence resulting in trust and distrust, respectively. Calibrated trust is the goal, but other psychological attitude states may occur that fall on these two axes. When high or low cognitive effort is inappropriately employed, other states such as suspicion, curiosity, complacency, and under reliance can occur. We talk about each of these constructs in relation to the model in Figure 5.

The trust process outlined above is an iterative rather than static process (Chiou and Lee, 2023; Lee and See, 2004). Many different cues from the AI/ML/XAI influence perceptions over time. As the user obtains more information about the AI/ML via use and observation, the trustworthiness perceptions are updated. These trustworthiness perceptions are then updated based on how the system responds. This is what Hoff and Bashir (2015) refer to as learned trust.

Importantly, learned trust is contextually specific in relation to trust. Trust in the system for one context may not be transferred to other contexts. Instead, different attitudes may arise depending on the time spent with the system, the risk involved with the task and the affective valence associated with the algorithm. This may be for a variety of reasons such as differential risk between scenarios. For example, Lyons et al. (2024) found failure of one autonomous system lowered the trust in that specific system but not to other systems with the same operating constraints. The learned trust is the information processing and updating of the mental models/schemas of the human in the human-AI/ML/XAI interaction over time. This creates a feedback loop where trustworthiness perceptions are updated over time and information becomes salient. However, other findings can emerge as well. System-wide trust literature notes that trust decay can bleed over into other independent systems (Keller and Rice, 2010). Across many studies in contexts ranging from in-flight aircraft services (e.g., Rice et al., 2016) to robotic swarm interactions (e.g., Capiola et al., 2024b), perceived perturbations of one system aspect can bring down trust toward another orthogonal system.

Trust and distrust

Workers attempt to make sense of the technology they utilize at the same time they are doing a given task (Muir, 1987). This links trust theoretically to trust toward AI/ML-enabled systems, and researchers have meta-analyzed data on trust-relevant constructs in human-AI/ML interactions (see Kaplan et al., 2023). As noted, XAI was developed because of the lack of transparency of black-box models like DNNs to help facilitate (dis)trust formation toward opaque systems (Sanneman and Shah, 2022). The construct of trust has been defined as an attitude (Lee and See, 2004) or willingness to be vulnerable to a referent (Mayer et al., 1995), and we adopt that definition for our model (see also Kohn et al., 2021). If the user has adequate transparency provided about the system's performance, purpose, and process, sufficient to outweigh the uncertainty and risk of the situation, they will trust the system appropriately. Individual differences, such as domain knowledge or task knowledge, may also play a role in this process as they provide information for perceiving the referent (Hoff and Bashir, 2015). The reason "appropriate" is used denotes the importance of veridical perceptions of the system (Lee and See, 2004), and this perhaps more important as referent machine systems leveraging AI/ML/XAI capabilities increase in ubiquity and opacity (Chiou and Lee, 2023; Schlicker et al., 2025). Indeed, if a user perceives the system to have low performance, purpose, and process, to the extent that their perceptions do not outweigh the perceived risk, then it is beneficial for them to *not* trust, i.e., distrust, the system (assuming their perceptions of the referent system are accurate). Trust should not be assumed to be beneficial in and of itself: what is important is that the user perceives the system accurately to form an intention to and ultimately rely on that system, again, what Lee and See (2004) note as calibrated trust, i.e., a correspondence between objective system capability and the user's subjective trust toward that system informed by their perceptions.

Appropriate trust calibration is not a function of one type of processing; instead, cognitive processing can occur at different stages such as during initial interactions or when new information is made salient (Stoltz and Lizardo, 2018; Tutić et al., 2024). Trust calibration and recalibration implies making appropriate use of the relevant information in the environment to make the assessment. For this reason, we place appropriate trust and distrust in the middle of the cognitive effort y-axis but on their respective sides of the valence x-axis. If new information is made salient, effortful processing increases to incorporate the information in the user's schema/attitude. However, the attitude of trust or distrust has been formed based on previous interactions which have resulted from cognitive effort and affective responses. Again, this is similar to what Hoff and Bashir (2015) term learned trust. This trust learned trust is a function of both system 1 and system 2 processing as the trust or distrust has been formed, aiding the user in processing information more efficiently, but new information that is not part of the schema will active system 2 processing. We note that recalibration is not illustrated in Figure 5 because it would depend on the previous perception. For example, if the user trusted the AI/ML/XAI and new information was provided, the user may be in the upper left quadrant.

> *Proposition 10: Calibrated trust and distrust are a function of moderate cognitive effort and their respective valence.*

Suspicion

Suspicion is defined as a state of increased cognitive activity, uncertainty, and perceiving machine system mal-intent (Bobko et al., 2014). Whereas trust and distrust are a willingness to rely or not rely on the system respectively, suspicion is the withholding of the evaluation due to uncertainty as to how the referent will behave. As such, the uncertainty leads to increased cognitive effort to determine whether to utilize the system. A system can also be perceived as suspicious because of an individual's tendency to be suspicious of machine systems (Calhoun et al., 2017), and contextual factors which facilitate suspicion may lead to distrust of a system with regards to its objective reliability (Bobko et al., 2014).

Little research has been conducted on suspicion toward AI/ML, despite recent calls for research on the construct in relation to AI/ML (Peters and Visser, 2023). Gay et al. (2017) found alerts from the AI/ML/XAI alone do not create suspicion. Instead, alerts facilitate information search through the users' increasing cognitive effort. Gay et al. note situational differences facilitate suspicion, such that negative information can lead to increased suspicion. This suspicion led to decreased user performance in their scenario. Similarly, Strang (2020) found suspicion is high in cyberspace operations. These results indicate it is the context, relative risk, and information gathered from the environment that facilitates suspicion. As such, suspicion is placed

in the upper right quadrant characterized by high cognitive effort and negative valence.

> *Proposition 11: High cognitive effort and negative valence will result in suspicion.*

Curiosity

State curiosity is a temporary, situational state of intrinsic motivation or desire to learn or explore (Kashdan and Roberts, 2004; Spielberger, 1979). State curiosity is externally triggered by aspects of the environment such as novelty, ambiguity, or gaps in knowledge. State curiosity involves both feelings (i.e., interest, excitement) and mental engagement with the referent (i.e., cognitive effort). Oudeyer et al. (2016) note that curiosity is formed when the agent's predictions are improving. Information gap theory (Loewenstein, 1994) theorizes state curiosity arises from an inconsistency or disparity between what is known and what is unknown. Curiosity is the drive for information through active intrinsic desire to obtain more information (de Abril and Kanai, 2018). Importantly, curiosity is driven by positive emotions in contrast to suspicion which is driven by negative emotions.

Research on curiosity in human-computer interaction research is relatively sparse. However, some researchers have developed models that personalize a user's curiosity appetite. Abbas and Niu (2019) found personalization of the system to users' openness to experience influenced more information gathering. Hoffman et al. (2023) also noted curiosity as an important factor in human-XAI interaction, noting that seeking information is driven by curiosity. Hoffman et al. note XAI should promote curiosity for increasing the accuracy of mental models, but no research to date has explored and demarcated the psychological "triggers" for curiosity. Still, it stands to reason that AI/ML/XAI can facilitate curiosity in the task. Researchers have found students interacting with ChatGPT led to more curiosity and creativity in the classroom (Essel et al., 2024). General curiosity can be triggered by a violation of expectations (Maheswaran and Chaiken, 1991), but it is the type of violation that distinguishes curiosity from suspicion. Curiosity may occur in less risky environments, when information obtained is not negative in the task, or when the context is relatively benign. Importantly, curiosity entails both positive valence (or at least a lack of mal-intent) and high cognitive effort for information seeking. However, some individuals may have a tendency to offer less cognitive effort (Petty and Cacioppo, 1986), and some situations may be so benign or so high in workload that exerting cognitive effort is not possible.

> *Proposition 12: High cognitive effort and positive valence will result in curiosity.*

Complacency

Complacency is defined as decreased vigilance, attentiveness, and situation awareness resulting in a user relying too heavily on a system (Parasuraman and Manzey, 2010). Complacency with the system can be due to many different aspects, but in the instance of over-trust, the user becomes overly confident that machine systems will handle task responsibilities (Lee and See, 2004). The lack of cognitive effort associated with complacency leads to less task vigilance. That is, complacency can lead to missed errors, reduced situation awareness, slower reaction times, increased risk of accidents, and skill degradation (Parasuraman and Manzey, 2010). Instances of this are easily accessible in the news with Tesla owners not paying attention to the car while it is in its autonomous mode (Shepardson and Sriram, 2024). Lack of system oversight has led to accidents, including deaths. However, the lack of oversight results in over use as the individual expects a positive outcome, which is why complacency is in the lower left quadrant of Figure 5.

One related construct to complacency that we depict in Figure 5 is algorithm appreciation (Logg et al., 2019), which describes users' tendency to rely on advice from algorithm referents compared to humans. Across several experiments, Logg et al. document cases of algorithm appreciation in tasks ranging from visual stimuli estimates and forecasting the popularity of content. Similar results are shown in foundational work in human-automation interaction (Dzindolet et al., 2002), where individuals demonstrate more positive attitudes toward machine decision support systems compared to humans *before* human-machine interaction commences.

Still, Logg et al. (2019) found appreciation for algorithms over humans when the latter forecast came from the participant themselves, or the participant was an expert in the forecasting context. Similarly, Dzindolet et al. (2002) found appreciation decreased after task interaction progressed, resulting in user *under* reliance on machine systems compared to humans when either was perceived to be imperfect.

> *Proposition 13: Low cognitive effort and positive valence will result in complacency.*

Under reliance

Under reliance on a system refers to insufficient or inadequate use of the machine despite its objective capabilities (Lee and See, 2004). Importantly, this disinclination to use a system is a function of a lack of cognitive effort as there is evidence that the system is reliable and beneficial (Parasuraman and Manzey, 2010). Still, a user may forego the potential benefits of relying on the system such as increased performance and decreased decision time due to the user's increased workload and potential fatigue. Thus, under reliance is comprised of lower cognitive effort exerted towards the system and an expectation of negative outcome, which is why under reliance is in the lower right corner of Figure 5. A good example of this can be found in the literature on algorithm aversion, which is the reluctance to use algorithms even when they demonstrate better accuracy and reliability than a human (Dietvorst et al., 2015). Negative experiences with other algorithms lead to the development of a negative bias toward algorithms when they are not perfect (Liu et al., 2023; Slovic et al., 2013). Additionally, negative emotions have been associated with less use of algorithms (Gogoll and Uhl, 2018; Prahl and Van Swol, 2017), leading to under reliance (i.e., algorithm aversion). This is driven by not only a bias or heuristic but also a function of one's emotional state or valence.

Gaube et al. (2024) found under reliance toward AI/ML was more harmful to performance than over-reliance. Moreover, they found that XAI reduced under reliance on AI/ML referents, especially when users were expected to classify difficult images, but under reliance on the system still led to lower task performance. Under reliance on machine systems can be the result of many different variables such as user lack of error tolerance (Dietvorst et al., 2015; Dzindolet et al., 2002), perceived low controllability of the AI/ML/XAI (Cheng and Chouldechova, 2023), low transparency (Schemmer et al., 2023), or poor mental models of the AI/ML by the user (Kaplan et al., 2023).

> *Proposition 14: Low cognitive effort and negative valence will result in under reliance.*

Personality can influence these attitudes through the mechanisms we described above as well as how users perceive the environment (Lazarus and Folkman, 1984). Personality is the lens through which humans view the world and process the information (McGuire, 1968). Different personality variables can influence the final attitude formation toward a given referent. However, the underlying processes of all the mechanisms through which personality can influence attitudes is beyond the scope of the current review. Interested readers are encouraged to review Albarracin and Shavitt (2018), Ajzen (2005), and Howe and Krosnick (2017).

Risk

The relevant uncertainty in utilizing an AI/ML/XAI algorithm represents an inherent risk in the task (Gulati, 1995). Risk is the uncertainty of the outcome and the relative vulnerability of relying on the system in the given context. Often times, this risk in an experimental task is instantiated as monetary payouts in the psychological trust literature (see Johnson and Mislin, 2011). However, this also has relevance to the use of AI/ML/XAI. The advent of XAI literature is in response to using black-box models in areas with high risk. The call for more transparent AI/ML is in direct response to the risk of utilizing these algorithms in parole decisions, autonomous vehicles, and other risky scenarios (Rudin, 2019). The role of situational risk is closely related to trust, suspicion, and over reliance. Risk augments how users trust and utilize the AI/ML by affecting how they perceive the fairness of the process/decision. Trust has also been viewed as a cognitive mechanism through which people process, interpret, and respond to informational risk. This risk perception varies by situation. A programmer developing an AI/ML algorithm to create a spam filter for email may have considerably less risk than a programmer utilizing AI/ML for customer payment processing systems. The risk inherent with each scenario will moderate how the user perceives the system (McComas, 2006; Thielmann and Hilbig, 2015), influencing their likelihood trusting that system given the tradeoff of accepting vulnerability toward that system in contexts of increased risk (Chiou and Lee, 2023; see also Kohn et al., 2021). Following the interpersonal trust literature (Mayer et al., 1995), contextual or perceived risk can also influence the processing resources users allocate toward making their decision to (dis)trust a machine referent (Kohn et al., 2021) including AI/ML/XAI (Chiou and Lee, 2023). Additionally, the risk inherent with each scenario can also establish which personality variables are activated in interactions with each algorithm. In the latter payment algorithm, there is considerably more risk; as such, personality variables such as risk aversion may play a larger role in the cognitive processing of information from the system. The former low risk scenario of the spam filter may not activate risk aversion because the risk is so low. Instead, personality variables such as complacency potential may be active in the cognitive processing because of the low inherent risk (Zhou et al., 2020).

Reliance behaviors

The cognitive processes (or lack thereof) described prior all lead to eventual decision-making which is referred to as reliance / compliance behaviors (Meyer, 2001; Meyer et al., 2014) or trust behaviors (Alarcon et al., 2021, 2023) in the Human Factors literature. As noted earlier, reliance is the actual dependence of the trustor on the trustee (i.e., delegation of a task). Much of the research has focused on attributing trust to reliance behaviors, such that appropriately trusting AI/ML means utilizing the system when it is accurate and disregarding the system when it is inaccurate. However, in real world applications pertinent in our theoretical model, the user will not know the actual state of the AI/ML/XAI decision. It is important that algorithms are designed so that users can most accurately trust them when it is applicable and appropriate.

An important issue in Psychology is that behaviors are not due to a single cognitive state. The Human Factors literature has noted several issues that may lead to a user's decision to utilize a system. Our discussion of attitudes in the previous section illustrates some of the different attitudes that can influence behaviors. Importantly, one additional reason for reliance behaviors may simply be user errors. A user may accidentally perform the correct behavior, either without knowing they were going to or because they "hit the wrong button," which happened to be correct. These human errors are often not accounted for in the literature, as it is assumed a user is performing the behavior on purpose. Instead, it may be that some of the behaviors are accounted for by simple mistakes.

Providing information about the system does not always lead to increased trust, nor are those increases always substantial. Atf and Lewis (2025) found XAI was only modestly correlated with trust assessments in their meta-analytic findings. Instead, performance aspects had the most robust relationship with trust. However, more information about the performance of a model, such as calibrated confidence intervals (Guo et al., 2017), may have implications for research such that models with more performance information are trusted more (Meyers et al., 2024; Harris et al., 2024). Explainability of the underlying processes can also lead to distrust in certain situations. The explainability can illustrate that the algorithm is using incorrect information in its decision-making. For example, there is the classic case of a CNN that has been trained to classify dogs and wolves might develop a bias due to the background of the images (Ribeiro et al., 2016a). Ribeiro et al. found this type of algorithm would classify an image as a wolf if the image had snow in the background. If a wolf without snow in the background is presented it is classified as a Husky, as the algorithm has not been trained on wolves that have no snow in the imagery. A model with XAI may illustrate the accuracy of the model is misplaced as it is not making a decision based on the relevant criteria, i.e., the animal of interest, but instead on other information such as the background. Additionally, too much information can lead to distrust, complacency, or under reliance. Mackay et al. (2019) found too much information led to over-trust in the system and decreased performance on a visual search task. The problem for display design in AI/ML/XAI is balancing the information provided with the information necessary to perform the task without overloading the human operator (see footnote 4; Young et al., 2015).

It is not just one construct that facilitates the use or disuse of a system. It is a combination of many variables that can influence reliance behaviors and the contextual moderators which increase or decrease the influence of each variable on reliance (or a lack thereof). Individual differences play a role in this aspect, too. For example, complacency potential can activate both overuse and underuse of the system. If the user is high in complacency potential and the system displays high performance, the user may overuse the system because they believe the system is not fallible (Shepardson and Sriram, 2024). Conversely, if the user is low in complacency and the system does not perform well, the user may

utilize another system or do the task themselves depending on the task. This is because the lack of need for cognition or complacency will lead the user to underestimate the system and abandon the AI/ML-enabled system.

Lastly, reliance behaviors have often been dichotomized in the literature (e.g., rely or not rely on the system). However, the relationship between the aforementioned attitudes and reliance is often not binary (Blanco, 2025). Research has demonstrated there is a richness of reliance behaviors. For example, Alarcon et al. (2023) found participants were willing to wager a little on a robot partner early in their interaction, but participants increased their wagers as time progressed because they trusted robot more over time. Researchers often classify reliance as "monitoring behaviors," but metrics such as eye-tracking (Sharafi et al., 2015) or time monitoring the algorithm are also not dichotomous. Indeed, the aforementioned attitudes can help to explain why a user may be spending more time monitoring an AI/ML/XAI. For example, a new algorithm that classifies images with more accurate confidence intervals (Guo et al., 2017) may lead to more time exploring how the algorithm made its decision. Here, time would be a continuous variable, not a binary variable. Additionally, this behavior could be caused by high cognitive effort and positive emotions such as curiosity pertaining to how the new algorithm outputs information.

Machine errors/mistakes and unintended consequences

Errors/mistakes and unintended consequences may have differential effects on trust. We differentiate between errors and unexpected outcomes. Errors are analogous to misses and false alarms in signal detection theory (Kay, 2013). In instances of errors, it can be something such as an incorrectly classifying an image, an incorrect forecasting decision, or misinformation provided by large language model (LLM). In contrast, unintended consequences are when something unforeseen occurs in the dataset or response. The example of the pregnant teenager example from Target also illustrates an unintended consequence. In that instance, the algorithm did perform its task well, but most would be reticent to send pregnancy related coupons to a minor. Lastly, LLMs can hallucinate or infringe on intellectual property. This can lead to unintended consequences such as plagiarism by LLMs. All of these can degrade user trust toward the system, but it remains to be seen if there is a difference between trust degradations due to errors and trust degradations due to unintended consequences.

Differences in performance, purpose and process

We note that performance perceptions are a necessary aspect of trust, at least after the first interaction. Much of the literature on DNNs has focused on the performance of the system, without discussing the underlying processes or purposes that drive DNNs (Minh et al., 2022).

The advent of XAI is a response to the lack of transparency in black-box models as they are used in more applications; as noted, governments and companies required more information on how the decision processes of the algorithms worked (European Union Act, 2024/1689; Air Force Doctrine Note, 2024). However, DNNs have been used in multiple contexts without transparency in process or purpose cues before, but the algorithms always displayed some performance information. Researchers note that no machine or system would be used without some kind of performance feedback; as such, performance is a necessary aspect across all temporal aspects of the trust process (Alarcon and Willis, 2023; Hoff and Bashir, 2015). As such, we theorize performance perceptions are a necessary condition for trust in the system. In scenarios where information about the system is either sparse or unknown, the user will leverage prior information either about machines in general (if the user has never interacted with AI/ML before) such as global beliefs and / or possible knowledge, skills, or abilities that the user holds (if the user is a domain expert) to make their initial trust assessment, which is really a strong belief without adequate, contextually constrained information.

Purpose perceptions are most relevant early in the trust process (Hoff and Bashir, 2015; Lee and See, 2004; Muir and Moray, 1996). Early interactions may be focused on performance and purpose perceptions and cues as the user lacks information about the system, such as reliability, dependability, or capabilities (Hoff and Bashir, 2015). The contextual nature of the system will be leveraged in the initial trust estimate as the user is unsure of how it will perform. As such, purpose perceptions along with global individual differences will have the strongest effect early on in interactions as there is not yet salient information about the system's performance.

Importantly, performance and purpose perceptions will have a strong relationship with each other, and quantitative data support this postulate (Alarcon et al., 2023; Capiola et al., in press). The reason a model was built will be highly correlated with performance in the context of the task.

Third, process perceptions may only be necessary after the AI/ML/XAI has made a mistake or its decision resulted in an unexpected consequence. As illustrated with the literature on AI/ML, many governments and companies were not concerned with the underlying processes of the DNNs until they started being utilized in high-risk scenarios and started to have errors or unexpected consequences (European Union Act, 2024/1689; Air Force Doctrine Note, 2024). We propose it is mainly when an algorithm makes a mistake or unintended consequence that users will be interested in the XAI or underlying processes (we note that developers will be interested in processes while training the model). This applies when training an algorithm, but this links back to the notion that training entails an inspection of results. In these situations, users need information to understand what went wrong. Recent experimental data on human-AI/ML interaction (Harris et al., 2024, 2025) shows this to be the case, providing fodder for future investigations with other AI/ML/XAI algorithms in different contexts. Moreover, a user's baseline expectations of a given AI/ML or XAI referent and their threshold for abandoning said systems should they err ought to shape their process perceptions differently, per the literature on perfect automation schema (Dzindolet et al., 2002). Thus, individual schemas for a system's function may shape user perceptions of how the system works given something unpredictable occurs.

Measurement

It is important to note the measurement of each of the variables in the proposed model. We note that Kohn et al. (2021) have covered the measurement of variables in the trust process extensively, but we highlight a few aspects here. Measurement of variables such as individual differences, trustworthiness perceptions, and attitudes about the system often leverage Likert-type scales. These scales are useful for collecting large amounts of data about the constructs of interest as they do not require much from the researcher. However, the specificity of the scales may impact their utility. As we have mentioned prior, cognitions and behaviors are based on several constructs. Solely using self-report scales may not facilitate the rich understanding that can come from the inclusion of many different methods (see Krausman et al., 2022). One such method is qualitative analyses.

Qualitative analyses provide a focus of meaning and context for the decisions of the user. These types of analyses can provide a rich and meaningful description of the thoughts and perceptions of the user (e.g., Meyers et al., 2025). However, a large drawback is the time and manpower needed to analyze and interpret the data.

Behavior is often the best indicator of behavioral trust. However, we note that much of the literature has focused on trust as a binary construct, with users or participants either trusting or not trusting in the scenario. Instead, trust may be more complex. Alarcon et al. (2023) utilized monetary risk to instantiate trust, with participants making a wager on how their robot or human partner would perform. Interestingly, when participants were allowed to choose their specific wager, they initially trusted a little with gradually increasing wagers as the partner demonstrated trustworthiness. This illustrates the idea that trust behaviors are not all-or-nothing, but rather iterative and build as trust develops. Similarly, as proxies of psychological trust can be assessed through a (lack of) monitoring behavior, physiological metrics such as eye-tracking can be used to determine how participants trust (i.e., do not monitor the AI/ML/XAI; see Krausman et al., 2022) the system over time. We propose that as trust develops there will be less oversight of the AI/ML/XAI.

Implications for research and practice

As noted earlier, the interpretation of the trustworthiness cues by the user is represented by the lines connecting the trustworthiness cues and trustworthiness perceptions. The gray boxes illustrate the degree to which features of the system are interpreted by the user, and the degree to which these cues facilitate trustworthiness perceptions (performance, purpose, and process) are moderated by individual differences in the human processor. For example, most AI/ML/XAI have focused on developer perceptions of the model rather than end user perceptions, but there are differences in between user's and developer's cognitive models and how they are formed (Ehsan et al., 2024). In depth discussions and experiments with both end users and developers can clarify what individuals need from the AI/ML/XAI within a given context. The various trustworthiness cues provide information about the state of the system (e.g., dependability, helpfulness, and comprehensibility) which are interpreted as subjective perceptions of the explanation data that is provided (Broniatowski and Broniatowski, 2021). The explanation types facilitate mental models comprising psychological perceptions of the algorithm's trustworthiness (Visser et al., 2023). However, these interpretations and cognitive schemas are not a one size fit all scenario, needing more attention paid to the specific user that is in mind. The output (e.g., XAI, beta weights, etc.) should provide a description of the stimulus, such as a data point or algorithm output, that can facilitate understanding within the context. In other words, the AI/ML/XAI should be able to communicate intentions and explain decision-making processes to the user (Boies et al., 2015; Paleja et al., 2021). These trustworthiness perceptions are cognitive evaluations by the user based on the relevant trustworthiness cues that are perceived, interpreted, and ultimately understood, assuming the system is appropriately transparent.

It is both the explanation type and proper display of the explanation type that makes the explainable data interpretable to the user according to information processing theory (Wickens and Carswell, 2012). This ease of interpretation of the explainable data leads to appropriate information processing by the user. Conversely, a poorly designed display of information can lead to a lack of or inappropriate information processing by the user. We note that humans prefer simpler explanations, and explanations should only grow in complexity when all the components of the explanation are highly accurate (Lombrozo and Carey, 2006; Lombrozo, 2007). Poor display design can influence information processing, resulting in an information overload hampering information processing (Bainbridge, 1983; Tocchetti and Brambilla, 2022). For example, feature importance metrics such as saliency measures have improved user's understanding of an AI/ML algorithm's decisions within an image classification task (Hase and Bansal, 2020). Research has noted that multiple signals simultaneously can confuse and disorient the user, especially if they represent different information (Hase and Bansal, 2020). If an XAI display were to provide multiple explanations simultaneously that are not highly correlated with each other (i.e., not redundant information), the user may be overloaded and not able to perceive relevant information (e.g., Capiola et al., 2024a). Indeed, Ngo (2025) found transparency had a curvilinear relationship with performance such that too much transparency inhibited performance. That is, if the sensory cues are all correlated and represent the same information, they may provide a failsafe with multiple cues about that information in some contexts but competition for finite resources in others (Wickens and Carswell, 2012).

Proper placement and display of the information is also necessary. Simply placing the information outside of the area of attention may influence the results of cognition and behavior, such that the user may not perceive the information or effectively process the information. For example, Ling et al. (2024) found confidence intervals were not important in decision-making with AI/ML; however, the authors displayed the confidence intervals as a graph at the top of the screen away from the target information. Eye tracking indicated participants did not utilize the data possibly because it was far from the focal point of the task and difficult to process. In contrast, research has demonstrated confidence intervals are utilized by participants in an image classification task when the confidence intervals are salient and near the target image (Alarcon et al., 2024; Harris et al., 2024). This illustrates our point further that poor display design can lead to a decreased focus on relevant information.

The interpretation of the display design information of the algorithm is a highly subjective interpretation. As we noted earlier

with the algorithm aversion studies, if the information is provided but the user lacks the ability to interpret the data, the explainability may be high (i.e., data is provided about the algorithm's decision-making), but the user's interpretability of the system is low (i.e., the user cannot make sense of the data). For example, the Dietvorst et al. (2015) research in which novice participants were required to interpret beta weights may have led to high explanation but low interpretation and thus misunderstanding.

Although we focused on the interpretability of white-box models with regressions for this example, this principle can also be described in terms of display design (e.g., an XAI displays what information an underlying AI/ML algorithm used to classify an image into a given category to a human user). Ehsan et al. (2024) found differences in between computer science students and the general public not only in their cognitive schemas that were formed, but also in the interpretation of the information that was provided by the AI/ML, indicating individual differences are key aspect of processing information.

Practical takeaways and remaining gaps

We offer several practical takeaways for researchers and designers alike to consider in their own work. First, individual differences matter in human-AI/ML/XAI interaction and are a cornerstone to human-centered design for effective human-machine interaction. Although AI/ML-enabled tools will be deployed to the general population or large organization (whether that be public or private), companies ought to consider at minimum what factors their intended users bring to bear when interacting with these technologies. Considering not just a target audience's knowledge, skills, and abilities with AI/ML-enabled systems, but also their general expectations of and thresholds for abandoning machines should be considered. Considering users' general tendency to engage, be curious, and think deeply about interactions with novel AI/ML-enabled systems will help shape design for effective use. Customized information display sensitive to users' individual differences may help them use a system appropriately (e.g., Vázquez-Ingelmo et al., 2019), and we encourage researchers to test this speculation by instantiating these strategies in emerging AI/ML-enabled systems coupled with different XAI applications.

Second, designers should consider what features of an AI/ML-enabled tool should be most perceptible to users. Our stance is an overabundance of information as to what the system did, how it did it, and why may be technically explainable but uninterpretable to some audiences (per individual differences above) and ultimately lead to misunderstanding. Simply: more information is not always better. Researchers should investigate specific features that are most effective at facilitating human understanding and what contextual factors and technological constraints of emerging AI/ML-enabled systems are necessary for the task. This approach has implications for *appropriately* explainable AI, which could be further leveraged by designers for product testing iteration. Though we suggested display customization to individuals immediately above, assuming an interface cannot be customized for *every* user, cataloging what information ought to be made most salient for users *in general* (e.g., performance information; Hoff and Bashir, 2015) is another prong of research we hope is explored in human-AI/ML/XAI research.

Finally, the behaviors that users engage in when leveraging AI/ML-enabled systems are determined by many factors. We contend that users' individual differences as well as their perceptions of a system's trustworthiness shape their willingness to use said systems. However, the appropriateness of system use can be guided by contextual risk, time constraints, and (sometimes) simple mistakes. As mentioned, the process by which behavior emerges is complex, comprising trust-relevant factors denoted and individual differences, but variability can be attributed to contextual aspects and emerging capabilities yet to be realized in AI/ML-enabled systems. We see our model's utility as a lens of analysis for researchers generating new questions about and designers imagining new interfaces for AI/ML-enabled systems and the XAI coupled to guide user engagement. We anticipate our model will be extended (and perhaps updated) as researchers investigate the edge-cases of human-AI/ML-XAI interaction and designers deploy novel AI/ML/XAI systems in the coming years.

Limitations

The current paper is not without limitations. First, the theoretical model we have established is based on research from literature across several different domains. Researchers have noted that trust is conceptualized differently across the computer, information technology, and social sciences. Indeed, this was the main drive for the current theoretical model, to elucidate the different constructs across fields. However, as the research cited in the current paper is based on research with different measures and conceptual ideas, it remains to be seen if the theoretical model withstands scrutiny across domains. Second, the theoretical model in the current paper was developed based upon previous research and as such is *ad hoc*. Future research should test the theoretical postulates with the explicit research hypotheses. This would allow for empirical testing of the postulates.

Finally, the present manuscript certainly did not cover each and every moderator of the trust process in human-machine interaction. Indeed, an anonymous reviewer noted details on the role of culture, norms, and group dynamics were missing from our model. These constructs were beyond the scope of the current work, but high-level cultural differences (e.g., Hoff and Bashir, 2015), norms for human-machine interaction (e.g., see footnote 4 and Cheng et al., 2016), and group dynamics in human-machine interaction (e.g., Demir et al., 2021) are discussed elsewhere. Future work can build upon our model, integrating these constructs and others in meaningful ways promoting further experimentation.

Conclusion

Trust toward complex machine systems like AI/ML and XAI is multiply determined. This paper outlines a litany of machine referents that may be differently trusted based on users' individual differences, contextual factors, and the interplay of these variables which shape trust and reliance. We do not propose that every variable relevant for trust toward AI/ML and XAI are mentioned here. Indeed, we assume with iterative research there will be others that arise and affect criteria, namely (in)appropriate trust and reliance / compliance. It is our hope that researchers leverage this theoretical approach for designing their experiments to test and expand findings on human-AI/ML/XAI interaction, as these interactions will only increase in frequency and stakes over the next decade and beyond.

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

GA: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. AC: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

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 author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Author disclaimer

The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, or the United States.

References

Abbas, F., and Niu, X. (2019). *One size does not fit all: modeling users' personal curiosity in recommender systems*. ArXivorg.

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Adadi, A., and Bouhoute, A. (2023). Explainable artificial intelligence for intelligent transportation systems: Are we there yet?. In A. Adadi and A. Bouhoute (Eds.) *Explainable Artificial Intelligence for Intelligent Transportation Systems* (pp.2–30), Taylor and Francis: Boca Raton, FL.

Air Force Doctrine Note (2024). AFDN 1-24, Artificial Intelligence

Ajzen, I. (2005). *Attitudes, personality and behaviour*. Berkshire: McGraw-Hill Education (UK).

Alarcon, G. M., Capiola, A., Hamdan, I. A., Lee, M. A., and Jessup, S. A. (2023). Differential biases in human-human versus human-robot interactions. *Appl. Ergon.* 106:103858. doi: 10.1016/j.apergo.2022.103858

Alarcon, G. M., Gamble, R., Jessup, S. A., Walter, C., Ryan, T. J., Wood, D. W., et al. (2017). Application of the heuristic-systematic model to computer code trustworthiness: the influence of reputation and transparency. *Cogent Psychol.* 4:1389640. doi: 10.1080/23311908.2017.1389640

Alarcon, G. M., Gibson, A. M., Jessup, S. A., and Capiola, A. (2021). Exploring the differential effects of trust violations in human-human and human-robot interactions. *Appl. Ergon.* 93:103350. doi: 10.1016/j.apergo.2020.103350

Alarcon, G. M., and Jessup, S. A. (2023). Propensity to trust and risk aversion: differential roles in the trust process. *J. Res. Pers.* 103:104349. doi: 10.1016/j.jrp.2023.104349

Alarcon, G. M., Jessup, S. A., Meyers, S. K., Johnson, D., and Bennette, W. D. (2024). Trustworthiness perceptions of machine learning algorithms: the influence of confidence intervals. In 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS) (pp. 1–6). IEEE.

Alarcon, G. M., Lyons, J. B., and Christensen, J. C. (2016). The effect of propensity to trust and familiarity on perceptions of trustworthiness over time. *Pers. Individ. Differ.* 94, 309–315. doi: 10.1016/j.paid.2016.01.031

Alarcon, G. M., Lyons, J. B., Christensen, J. C., Klosterman, S. L., Bowers, M. A., Ryan, T. J., et al. (2018). The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads. *Behav. Res. Methods* 50, 1906–1920. doi: 10.3758/s13428-017-0959-6

Alarcon, G. M., and Ryan, T. J. (2018). Trustworthiness perceptions of computer code: a heuristic-systematic processing model. Proceedings of the 51st Hawaii International Conference on System Science, pp. 5384–5393. Available online at: http://hdl.handle.net/10125/50560

Alarcon, G., and Willis, S. (2023). Explaining explainable artificial intelligence: an integrative model of objective and subjective influences on XAI. Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 1095–1104. Available online at: https://hdl.handle.net/10125/102764

Albarracin, D., and Shavitt, S. (2018). Attitudes and attitude change. *Annu. Rev. Psychol.* 69, 299–327. doi: 10.1146/annurev-psych-122216-011911

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99:101805. doi: 10.1016/j.inffus.2023.101805

Angelov, P., and Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Netw.* 130, 185–194. doi: 10.1016/j.neunet.2020.07.010

Atf, Z., and Lewis, P. R. (2025). Is trust correlated with explainability in AI? A meta-analysis. *IEEE Trans. Technol. Soc.*, 1–8. doi: 10.1109/TTS.2025.3558448

Audi, R. (1999). *The Cambridge dictionary of philosophy, vol. 2*. Cambridge: Cambridge University Press.

Bainbridge, L. (1983). "Ironies of automation" in *Analysis, design and evaluation of man– machine systems* (Science Direct, German: Pergamon), 129–135.

Baird, A., and Maruping, L. M. (2021). The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MIS Q.* 45, 315–341. doi: 10.25300/MISQ/2021/15882

Barreto Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Berger, B., Adam, M., Rühr, A., and Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Bus. Inf. Syst. Eng.* 63, 55–68. doi: 10.1007/s12599-020-00678-5

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. Columbus, Oh: McGraw-Hill.

Berlyne, D. E. (1978). Curiosity and learning. *Motiv. Emot.* 2, 97–175. doi: 10.1007/BF00993037

Biros, D. P., Daly, M., and Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decis. Negot.* 13, 173–189. doi: 10.1023/B:GRUP.0000021840.85686.57

Blanco, S. (2025). Human trust in AI: a relationship beyond reliance. *AI Ethics* 5, 4167–4180. doi: 10.1007/s43681-025-00690-z

Bobko, P., Barelka, A. J., Hirshfield, L. M., and Lyons, J. B. (2014). Invited article: The construct of suspicion and how it can benefit theories and models in organizational science. *J. Bus. Psychol.* 29, 335–342. doi: 10.1007/s10869-014-9360-y

Boies, K., Fiset, J., and Gill, H. (2015). Communication and trust are key: unlocking the relationship between leadership and team performance and creativity. *Leadersh. Q.* 26, 1080–1094. doi: 10.1016/j.leaqua.2015.07.007

Broniatowski, D. A., and Broniatowski, D. A. (2021). *Psychological foundations of explainability and interpretability in artificial intelligence*. Washington, D. C.: US Department of Commerce, National Institute of Standards and Technology.

Bunghez, C., De Houwer, J., Rusu, A., and Sava, F. A. (2024). Exploring the association between neuroticism and negativity bias in evaluative counterconditioning. *Learn. Motiv.* 88:102039. doi: 10.1016/j.lmot.2024.102039

Burrell, J. (2016). How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc.* 3, 1–12. doi: 10.1177/2053951715622512

Cacioppo, J. T., and Petty, R. E. (1982). The need for cognition. *J. Pers. Soc. Psychol.* 42, 116–131. doi: 10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., and Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychol. Bull.* 119, 197–253. doi: 10.1037/0033-2909.119.2.197

Calhoun, C., Bobko, P., Schuelke, M., Jessup, S., Ryan, T., Walter, C., et al. (2017). *Suspicion, trust, and automation*. Dayton, Oh: SRA International Inc. Publication Report No. AFRL-RH-WP-TR-2017-0002.

Capiola, A., Fox, E. L., Bowers, G., Harris, K. N., and Woods, S. (2024a). Investigating interfaces that convey team efficiency. Proceedings of the 4th International Conference on Human-Machine Systems, Toronto, ON, Canada

Capiola, A., Hamdan, I. A., Lyons, J. B., Lewis, M., Alarcon, G. M., and Sycara, K. (2024b). The effect of asset degradation on trust in swarms: a reexamination of system-wide trust in human-swarm interaction. *Hum. Factors* 66, 1475–1489. doi: 10.1177/00187208221145261

Capiola, A., Harris, K. N., Alarcon, G. M., Johnson, D., Jessup, S. A., Willis, S. M., et al. (in press). Are you sure about that?" The effects of calibrated classification model task accuracy and confidence on trustworthiness, trust, and performance. Applied Ergonomics.

Capiola, A., Lyons, J. B., Harris, K. N., aldin Hamdan, I., Kailas, S., and Sycara, K. (2023). "Do what you say?" the combined effects of framed social intent and autonomous agent behavior on the trust process. *Comput. Human Behav.* 149:107966. doi: 10.1016/j.chb.2023.107966

Capogrosso, A., Treffers, T., and Welpe, I. (2025). Interaction context is key: a meta-analysis of experimental evidence on interventions against algorithm aversion. Proceedings of the 58th Hawaii International Conference on System Sciences (pp. 600–609).

Casey, L. J., and Vogel, M. D. (2019). Preparing for the next generation: profiles of millennial city managers and their approach to the job. *State Local Gover. Rev.* 51, 122–133. doi: 10.1177/0160323X19889094

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* 39, 752–766. doi: 10.1037/0022-3514.39.5.752

Cheng, L., and Chouldechova, A. (2023). Overcoming algorithm aversion: a comparison between process and outcome control. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–27).

Cheng, X., Fu, S., Sun, J., Han, Y., Shen, J., and Zarifis, A. (2016). Investigating individual trust in semi-virtual collaboration of multicultural and unicultural teams. *Comput. Human Behav.* 62, 267–276. doi: 10.1016/j.chb.2016.03.093

Chiou, E. K., and Lee, J. D. (2023). Trusting automation: designing for responsivity and resilience. *Hum. Factors* 65, 137–165. doi: 10.1177/00187208211009995

Colin, J., Fel, T., Cadène, R., and Serre, T. (2022). What I cannot predict, I do not understand: a human-centered evaluation framework for explainability methods. *Adv. Neural Inf. Process. Syst.* 35, 2832–2845

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909–927. doi: 10.1037/0021-9010.92.4.909

Costa, P. T. Jr., and McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *J. Personal. Disord.* 6, 343–359. doi: 10.1521/pedi.1992.6.4.343

de Abril, I. M., and Kanai, R. (2018). Curiosity-driven reinforcement learning with homeostatic regulation. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1–6 IEEE

Demir, M., McNeese, N. J., Gorman, J. C., Cooke, N. J., Myers, C. W., and Grimm, D. A. (2021). Exploration of teammate trust and interaction dynamics in human-autonomy teaming. *IEEE Trans. Hum.-Mach. Syst.* 51, 696–705. doi: 10.1109/THMS.2021.3115058

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126. doi: 10.1037/xge0000033

Drake, J. (2018). *Introduction to logic*. Waltham Abbey Esssex: EP Tech Press, 160–161.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2021). Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* 57:101994. doi: 10.1016/j.ijinfomgt.2019.08.002

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Hum. Factors* 44, 79–94. doi: 10.1518/0018720024494856

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I. H., Muller, M., et al. (2024). The who in XAI: how AI background shapes perceptions of AI explanations. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1–32).

Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Comput. Hum. Behav.* 140, 107574.

Essel, H. B., Vlachopoulos, D., Essuman, A. B., and Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: receiving instant responses from AI-based conversational large language models (LLMs). *Comput. Educ. Artif. Intell.* 6:100198. doi: 10.1016/j.caeai.2023.100198

European Union Act (2024/1689). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. *Am. Psychol.* 56, 218–226. doi: 10.1037//0003-066x.56.3.218

Gaube, S., Jussupow, E., Kokje, E., Khan, J., Bondi-Kelly, E., Schicho, A., et al. (2024). Underreliance harms human-AI collaboration more than overreliance in medical imaging. OSF Preprint. Available online at: https://safe.menlosecurity.com/doc/docview/viewer/docN209E3F6213C46ab6ba53a8a5d1143e42ae1e34ff9a682800c2bd0c46dd7f6c1e7684eaea2ad9

Gay, C., Horowitz, B., Elshaw, J., Bobko, P., and Kim, I. (2017). Operator suspicion and decision responses to cyber-attacks on unmanned ground vehicle systems. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 61, No. 1 (pp. 226–230). Los Angeles, CA: SAGE Publications.

Gibson, A. M., Capiola, A., Alarcon, G. M., Lee, M. A., Jessup, S. A., and Hamdan, I. A. (2023). Construction and validation of an updated perfect automation schema (uPAS) scale. *Theor. Issues Ergon. Sci.* 24, 241–266. doi: 10.1080/1463922X.2022.2081375

Gogoll, J., and Uhl, M. (2018). Rage against the machine: automation in the moral domain. *J. Behav. Exp. Econ.* 74, 97–103. doi: 10.1016/j.socec.2018.04.003

Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* 38, 50–57. doi: 10.1609/aimag.v38i3.2741

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 1–42. doi: 10.1145/3236009

Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Acad. Manag. J.* 38, 85–112. doi: 10.2307/256729

Gunning, D., and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* 40, 44–58. doi: 10.1609/aimag.v40i2.2850

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017) On calibration of modern neural networks. Proceedings of the International Conference on Machine Learning (pp. 1321–1330). Available online at: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

Harris, K. N., Capiola, A., Alarcon, G. M., Johnson, D., Jessup, S., Willis, S., et al. (2024). Exploring the effects of machine learning algorithms of varying transparency on performance outcomes. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (p. 10711813241260350). Los Angeles, CA: SAGE Publications.

Harris, K. N., Capiola, A., Johnson, D., Alarcon, G. M., Jessup, S. A., Willis, S., et al. (2025). Investigating the effects of classification model error type on trust-relevant criteria in a human-machine learning interaction task. Proceedings of the 58th Hawaii International Conference on System Sciences. (pp. 302–311). Available online at: https://hdl.handle.net/10125/108873

Hase, P., and Bansal, M. (2020). Evaluating explainable AI: which algorithmic explanations help users predict model behavior? *arXiv* [Preprint]. *arXiv:2005.01831*. doi: 10.48550/arXiv.2005.01831

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., et al. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* 16, 45–74. doi: 10.1007/s12559-023-10179-8

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., et al. (2019). Scaling out-of-distribution detection for real-world settings. *arXiv* [Preprint]. *arXiv:1911.11132*. doi: 10.48550/arXiv.1911.11132

Herm, L. V., Heinrich, K., Wanner, J., and Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *Int. J. Inf. Manag.* 69:102538. doi: 10.1016/j.ijinfomgt.2022.102538

Herm, L. V., Wanner, J., Seubert, F., and Janiesch, C. (2021) I don't get it, but it seems valid! The connection between explainability and comprehensibility. In (X) AI Research European Conference on Information Systems, Virtual Conference

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* 5:1096257. doi: 10.3389/fcomp.2023.1096257

Holstein, T., Dodig-Crnkovic, G., and Pelliccione, P. (2018). Ethical and social aspects of self-driving cars. *arXiv* [Preprint]. *arXiv:1802.04103*.

Hoofnagle, C. J., Van Der Sloot, B., and Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. *Inf. Commun. Technol. Law* 28, 65–98. doi: 10.1080/13600834.2019.1573501

Hoong, A. L. S., Thi, L. S., and Lin, M.-H. (2017). "Affective technology acceptance model: extending technology acceptance model with positive and negative affect" in *Knowledge management strategies and applications*. eds. A. Kobir, C. Yuliang, M. Mohiuddin and N. Halilem, Rijeka, Croatia 147–165.

Howe, L. C., and Krosnick, J. A. (2017). Attitude strength. *Annu. Rev. Psychol.* 68, 327–351. doi: 10.1146/annurev-psych-122414-033600

Hradecky, D., Kennell, J., Cai, W., and Davidson, R. (2022). Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe. *Int. J. Inf. Manag.* 65:102497. doi: 10.1016/j.ijinfomgt.2022.102497

Hübner, D. (2021). Two kinds of discrimination in AI-based penal decision-making. *ACM SIGKDD Explor. Newsl.* 23, 4–13. doi: 10.1145/3468507.3468510

Hu, W. C., Wu, H. T., Cho, H. H., and Tseng, F. H. (2020). Optimal route planning system for logistics vehicles based on artificial intelligence. *J. Internet Technol.* 21, 757–764.

Isen, A. M., and Geva, N. (1987). The influence of positive affect on acceptable level of risk: the person with a large canoe has a large worry. *Organ. Behav. Hum. Decis. Process.* 39, 145–154. doi: 10.1016/0749-5978(87)90034-3

Jessup, S. A., Alarcon, G. M., Meyers, S. K., Harris, K. N., Capiola, A., and Noblick, J. (2025). Jazz hands and jitters: exploring valence and arousal dimensions with multidimensional scaling techniques. *Pers. Individ. Differ.* 233:article 112930.

Jessup, S. A., Alarcon, G. M., Willis, S. M., and Lee, M. A. (2024). A closer look at how experience, task domain, and self-confidence influence reliance towards algorithms. *Appl. Ergon.* 121:104363. doi: 10.1016/j.apergo.2024.104363

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., and Capiola, A. (2019) The measurement of the propensity to trust automation. In Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21 (pp. 476–489) Springer International Publishing

Jessup, S. J., Gibson, A. M., Capiola, A., Alarcon, G. M., and Borders, M. (2020). Investigating the effect of trust manipulations on affect over time in human-human versus human-robot interactions. Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 553–562.

Jessup, S., Willis, S. M., and Alarcon, G. (2023). Extending the affective technology acceptance model to human-robot interactions: a multi-method perspective. Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 491–500. doi: 10.1016/j.paid.2024.112930

Johnson, N. D., and Mislin, A. A. (2011). Trust games: a meta-analysis. *J. Econ. Psychol.* 32, 865–889. doi: 10.1016/j.joep.2011.05.007

Jones, S. L., and Shah, P. P. (2016). Diagnosing the locus of trust: a temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. *J. Appl. Psychol.* 101, 392–414. doi: 10.1037/apl0000041

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2023). Trust in artificial intelligence: meta-analytic findings. *Hum. Factors* 65, 337–359. doi: 10.1177/00187208211013988

Kaplan, A. D., Kessler, T. T., and Hancock, P. A. (2020). How trust is defined and its use in human-human and human-machine interaction. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 64, No. 1 (pp. 1150–1154). Los Angeles, CA, SAGE Publications.

Kashdan, T. B., and Roberts, J. E. (2004). Trait and state curiosity in the genesis of intimacy: differentiation from related constructs. *J. Soc. Clin. Psychol.* 23, 792–816. doi: 10.1521/jscp.23.6.792.54800

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., and Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW) (pp. 169–175). IEEE.

Kay, S. M. (2013). *Fundamentals of statistical signal processing: Practical algorithm development (vol. 3)*. Wesford, MA: Pearson Education.

Kazén, M., Kuhl, J., and Quirin, M. (2015). Personality interacts with implicit affect to predict performance in analytic versus holistic processing. *J. Pers.* 83, 251–261. doi: 10.1111/jopy.12100

Keane, M. T., and Kenny, E. M. (2019). How case-based reasoning explains neural networks: a theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In Case-Based Reasoning Research and Development: 27th International Conference, ICCBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings 27 (pp. 155–171). Springer International Publishing.

Keller, D., and Rice, S. (2010). System-wide versus component-specific trust using multiple aids. *J. Gen. Psychol.* 137, 114–128. doi: 10.1080/00221300903266713

Kim, A., Yang, M., and Zhang, J. (2023). When algorithms err: differential impact of early vs. late errors on users' reliance on algorithms. *ACM Trans. Comput.-Hum. Interact.* 30, 1–36. doi: 10.1145/3557889

Klein, G., Moon, B., and Hoffman, R. R. (2006). Making sense of sensemaking 1: alternative perspectives. *IEEE Intell. Syst.* 21, 70–73. doi: 10.1109/MIS.2006.75

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychol. Bull.* 110, 499–519. doi: 10.1037/0033-2909.110.3.499

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C., and Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. *Front. Psychol.* 12:604977. doi: 10.3389/fpsyg.2021.604977

Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A. L., Fitzhugh, S. M., et al. (2022). Trust measurement in human-autonomy teams: development of a conceptual toolkit. *ACM Trans. Hum.-Robot Interact.* 11, 1–58. doi: 10.1145/3530874

Lazarus, R. S., and Folkman, S. (1984). *Stress, appraisal, and coping*. New York, NY: Springer.

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50_30392

Lee, M. A., Alarcon, G. M., and Capiola, A. (2022). "I think you're trustworthy, need I say more?" the factor structure and practicalities of trustworthiness assessment. *Front. Psychol.* 13:797443. doi: 10.3389/fpsyg.2022.797443

Li, B., Xu, Z., Hong, N., and Hussain, A. (2022). A bibliometric study and science mapping research of intelligent decision. *Cogn. Comput.* 14, 989–1008. doi: 10.1007/s12559-022-09993-3

Ling, S., Zhang, Y., and Du, N. (2024). More is not always better: impacts of AI-generated confidence and explanations in human–automation interaction. *Hum. Factors* 66, 2606–2620. doi: 10.1177/00187208241234810

Liu, M., Tang, X., Xia, S., Zhang, S., Zhu, Y., and Meng, Q. (2023). Algorithm aversion: evidence from ridesharing drivers. *Manag. Sci.* doi: 10.1287/mnsc.2022.02475

Li, Y., Kobsa, A., Knijnenburg, B. P., and Nguyen, M. C. (2017). Cross-cultural privacy prediction. *Proc. Priv. Enhanc. Technol.* doi: 10.1515/popets-2017-0019

Loewenstein, G. (1994). The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* 116, 75–98. doi: 10.1037/0033-2909.116.1.75

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257. doi: 10.1016/j.cogpsych.2006.09.006

Lombrozo, T., and Carey, S. (2006). Functional explanation and the function of explanation. *Cognition* 99, 167–204. doi: 10.1016/j.cognition.2004.12.009

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., et al. (2024). Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* 106:102301. doi: 10.1016/j.inffus.2024.102301

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv* [Preprint]. *arXiv:1705.07874*.

Lyons, J. B. (2013). Being transparent about transparency: a model for human-robot interaction. In 2013 AAAI Spring Symposium Series.

Lyons, J. B., Hamdan, I., and Vo, T. (2023). Explanations and trust: what happens to trust when a robot partner does something unexpected? *Comput. Human Behav.* 138:Article 107473. doi: 10.1016/j.chb.2022.107473

Lyons, J. B., Mator, J. D., Orr, T., Alarcon, G. M., and Barrera, K. (2024). Is the pull-down effect overstated? An examination of trust propagation among fighter pilots in a high-fidelity simulation. *J. Cogn. Eng. Decis. Making* 18, 99–113. doi: 10.1177/15553434231225909

Lyons, J. B., Sycara, K., Lewis, M., and Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. *Front. Psychol.* 12, 589585.

Mackay, A., Fortes, I., Santos, C., Machado, D., Barbosa, P., Boas, V. V., et al. (2019) The impact of autonomous vehicles' active feedback on trust International Conference on Applied Human Factors and Ergonomics 342–352 Cham Springer International Publishing

Maheswaran, D., and Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: effect of incongruent information on processing and judgment. *J. Pers. Soc. Psychol.* 61, 13–25. doi: 10.1037/0022-3514.61.1.13

Mahmud, M., Kaiser, M. S., McGinnity, T. M., and Hussain, A. (2021). Deep learning in mining biological data. *Cogn. Comput.* 13, 1–33. doi: 10.1007/s12559-020-09773-x

Maroto-Gómez, M., Alonso-Martín, F., Malfaz, M., Castro-González, Á., Castillo, J. C., and Salichs, M. Á. (2023). A systematic literature review of decision-making and control systems for autonomous and social robots. *Int. J. Soc. Robot.* 15, 745–789. doi: 10.1007/s12369-023-00977-3

Matthews, G. (2008). "Personality and information processing: a cognitive-adaptive theory" in *Handbook of personality theory and assessment*, Los Angeles, CA vol. 1, 56–79.

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

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manage. J.* 38, 24–59.

McComas, K. A. (2006). Defining moments in risk communication research: 1996–2005. *J. Health Commun.* 11, 75–91. doi: 10.1080/10810730500461091

McGuire, W. J. (1968). "Personality and attitude change: an information-processing theory" in *Psychological foundations of attitudes*. eds. A. Greenwald, T. Brock and T. Ostrom (New York: AcademicPress).

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Hum. Factors* 53, 356–370. doi: 10.1177/0018720811411912

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., et al. (2019). Automation-induced complacency potential: development and validation of a new scale. *Front. Psychol.* 10:225. doi: 10.3389/fpsyg.2019.00225

Merritt, S. M., Unnerstall, J. L., Lee, D., and Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Hum. Factors* 57, 740–753. doi: 10.1177/0018720815581247

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Hum. Factors* 43, 563–572. doi: 10.1518/001872001775870395

Meyer, J., Wiczorek, R., and Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Hum. Factors* 56, 840–849. doi: 10.1177/0018720813512865

Meyers, S., Capiola, A., Alarcon, G. M., and Bennette, W. (2024). Transparency and trustworthiness: exploring human-machine interaction in an image classification task. In 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS) (pp. 1–6). IEEE.

Meyers, S., Murry, P., Jessup, S., Alarcon, G., and Harris, K. (2025). Exploring user evaluations of machine learning models: A qualitative study on the impact of confidence intervals. *Hawaii International Conference on System Sciences*, 841–851.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artif. Intell. Rev.* 55, 3503–3568. doi: 10.1007/s10462-021-10088-y

Mitchell, D. J., Edward Russo, J., and Pennington, N. (1989). Back to the future: temporal perspective in the explanation of events. *J. Behav. Decis. Mak.* 2, 25–38. doi: 10.1002/bdm.3960020103

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man-Mach. Stud.* 27, 527–539.

Muir, B. M., and Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Müller, R., Thoß, M., Ullrich, J., Seitz, S., and Knoll, C. (2024). Interpretability is in the eye of the beholder: human versus artificial classification of image segments generated by humans versus XAI. *Int. J. Hum.-Comput. Interact.* 41, 1–23. doi: 10.1080/10447318.2024.2323263

Ngo, V. M. (2025). Balancing AI transparency: trust, certainty, and adoption. *Inf. Dev.*:02666669251346124. doi: 10.1177/02666669251346124

Oprescu, A. M., Miro-Amarante, G., García-Díaz, L., Beltrán, L. M., Rey, V. E., and Romero-Ternero, M. (2020). Artificial intelligence in pregnancy: a scoping review. *IEEE Access* 8, 181450–181484. doi: 10.1109/ACCESS.2020.3028333

Oudeyer, P. Y., Gottlieb, J., and Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: theory and applications in educational technologies. *Prog Brain Res* 229, 257–284. doi: 10.1016/bs.pbr.2016.05.005

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. *Adv. Neural Inf. Process. Syst.* 34, 610–623.

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* 52, 381–410. doi: 10.1177/0018720810376055

Peters, T. M., and Visser, R. W. (2023). The importance of distrust in AI World Conference on Explainable Artificial Intelligence 301–317 Cham Springer Nature Switzerland

Petty, R. E., and Cacioppo, J. T. (1986). "The elaboration likelihood model of persuasion" in *Advances in experimental social psychology, Vol. 19*. ed. L. Berkowitz (New York, NY: Academic Press), 123–205.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–52)

Prahl, A., and Van Swol, L. (2017). Understanding algorithm aversion: when is advice from automation discounted? *J. Forecast.* 36, 691–702. doi: 10.1002/for.2464

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv* [Preprint]. *arXiv:1606.05386*. doi: 10.48550/arXiv.1606.05386

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

Rice, S., Winter, S. R., Deaton, J. E., and Cremer, I. (2016). What are the predictors of system-wide trust loss in transportation automation? *J. Aviat. Technol. Eng.* 6, 1–8. doi: 10.7771/2159-6670.1120

Roberts, B. W., and Yoon, H. J. (2022). Personality psychology. *Annu. Rev. Psychol.* 73, 489–516. doi: 10.1146/annurev-psych-020821-114927

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8, 42200–42216. doi: 10.1109/ACCESS.2020.2976199

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K. R. (2021). Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109, 247–278. doi: 10.1109/JPROC.2021.3060483

Sanneman, L., and Shah, J. A. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *Int. J. Hum.-Comput. Interact.* 38, 1772–1788. doi: 10.1080/10447318.2022.2081282

Schaefer, K. E., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 58, 377–400. doi: 10.1177/0018720816634228

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., and Satzger, G. (2023) Appropriate reliance on AI advice: conceptualization and the effect of explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 410–422)

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., and Langer, M. (2025). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Comput. Human Behav.* 170:108671. doi: 10.1016/j.chb.2025.108671

Sharafi, Z., Soh, Z., and Guéhéneuc, Y. G. (2015). A systematic literature review on the usage of eye-tracking in software engineering. *Inf. Softw. Technol.* 67, 79–107. doi: 10.1016/j.infsof.2015.06.008

Shepardson, D., and Sriram, A. (2024). US probes tesla's full self-driving software in 2.4 mln cars after fatal crash. Reuters. Available online at: https://www.reuters.com/business/autos-transportation/nhtsa-opens-probe-into-24-mln-tesla-vehicles-over-full-self-driving-collisions-2024-10-18/

Shepherd, J. (2022). Conscious cognitive effort in cognitive control. *Wiley Interdiscip. Rev. Cogn. Sci.* 14:e1629. doi: 10.1002/wcs.1629

Siau, K., and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus. Technol. J.* 31, 47–53.

Singh, A. K., Khan, M. A., Singh, A., and Maheshwari, A. (2021). Color vision in civil aviation. *Indian J. Ophthalmol.* 69, 1032–1037. doi: 10.4103/ijo.IJO_2252_20

Singh, I. L., Molloy, R., and Parasuraman, R. (1993). Automation-induced "complacency": development of the complacency-potential rating scale. *Int. J. Aviat. Psychol.* 3, 111–122. doi: 10.1207/s15327108ijap0302_2

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2013). "Risk as analysis and risk as feelings: some thoughts about affect, reason, risk and rationality" in *The feeling of risk* (Routledge, Oxfordshire, UK: Routledge), 21–36.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 2239–2250.

Spielberger, C. D. (1979) Preliminary Manual for the State–Trait Personality Inventory (STPI). Unpublished manuscript, University of South Florida, Tampa

Stevens, A. F., and Stetson, P. (2023). Theory of trust and acceptance of artificial intelligence technology (TrAAIT): an instrument to assess clinician trust and acceptance of artificial intelligence. *J. Biomed. Inform.* 148:104550. doi: 10.1016/j.jbi.2023.104550

Stoltz, D. S., and Lizardo, O. (2018). Deliberate trust and intuitive faith: a dual-process model of reliance. *J. Theory Soc. Behav.* 48, 230–250. doi: 10.1111/jtsb.12160

Strang, M. G. (2020) Recognizing Potential Cyberspace Warriors through the Use of Suspicion Propensity Index

Suh, A., Hurley, I., Smith, N., and Siu, H. C. (2025). Fewer than 1% of explainable ai papers validate explainability with humans. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1–7).

Szalma, J. L. (2009). Individual differences in human-technology interaction: incorporating variation in human characteristics into human factors and ergonomics research and design. *Theor. Issues Ergon. Sci.* 10, 381–397. doi: 10.1080/14639220902893613

Szalma, J. L. (2008). "Individual differences in stress reaction" in *Performance under stress*. eds. P. A. Hancock and J. L. Szalma (Hampshire, UK: Ashgate), 323–357.

Szalma, J. L., and Taylor, G. S. (2011). Individual differences in response to automation: the five factor model of personality. *J. Exp. Psychol. Appl.* 17, 71–96. doi: 10.1037/a0024170

Thielmann, I., and Hilbig, B. E. (2015). Trust: an integrative review from a person–situation perspective. *Rev. Gen. Psychol.* 19, 249–277.

Tocchetti, A., and Brambilla, M. (2022). The role of human knowledge in explainable AI. *Data* 7:93. doi: 10.3390/data7070093

Tutić, A., Grehl, S., and Liebe, U. (2024). A dual-process perspective on the relationship between implicit attitudes and discriminatory behaviour. *Eur. Sociol. Rev.* 40, 672–685. doi: 10.1093/esr/jcad067

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum.-Comput. Interact.* 7, 1–38. doi: 10.1145/3579605

Vázquez-Ingelmo, A., García-Peñalvo, F. J., and Therón, R. (2019). Tailored information dashboards: a systematic mapping of the literature. In Proceedings of the XX International Conference on Human Computer Interaction (pp. 1–8).

Vilone, G., and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* 76, 89–106. doi: 10.1016/j.inffus.2021.05.009

Visser, R., Peters, T. M., Scharlau, I., and Hammer, B. (2023). Trust, distrust, and appropriate reliance in (X) AI: a survey of empirical evaluation of user trust. *arXiv* [Preprint]. *arXiv:2312.02034*. doi: 10.48550/arXiv.2312.02034

Waris, O., Soveri, A., Lukasik, K. M., Lehtonen, M., and Laine, M. (2018). Working memory and the big five. *Pers. Individ. Differ.* 130, 26–35. doi: 10.1016/j.paid.2018.03.027

Westbrook, A., and Braver, T. S. (2015). Cognitive effort: a neuroeconomic approach. *Cogn. Affect. Behav. Neurosci.* 15, 395–415. doi: 10.3758/s13415-015-0334-y

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* 3, 159–177. doi: 10.1080/14639220210123806

Wickens, C. D., and Carswell, C. M. (2012). "Information processing" in *Handbook of human factors and ergonomics*. ed. G. Salvendy (Routledge, Oxfordshire, UK: Wiley), 114–158.

Wickens, C. D., and Hollands, J. (2000). *Engineering psychology and human performance. 3rd* Edn. Upper Saddle River, NJ: Prentice-Hall.

Xing, J. (2006). *Color and visual factors in ATC displays* (No. DOTFAAAM0615).

Young, M. S., Brookhuis, K. A., Wickens, C. D., and Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics* 58, 1–17. doi: 10.1080/00140139.2014.956151

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: a survey on methods and metrics. *Electronics* 10:593. doi: 10.3390/electronics10050593

Zhou, J., Luo, S., and Chen, F. (2020). Effects of personality traits on user trust in human–machine collaborations. *J. Multimodal User Interfaces* 14, 387–400. doi: 10.1007/s12193-020-00329-9