frontiers | Research Topics

Disinformation countermeasures and artificial intelligence

Edited by
Oleksandr Makarenko, Paul Vines, George Cybenko
and Ludmilla Huntsman

Published in
Frontiers in Artificial Intelligence
Frontiers in Political Science
Frontiers in Big Data

ISSN 1664-8714
ISBN 978-2-8325-7228-3
DOI 10.3389/978-2-8325-7228-3

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook 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.

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

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

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

Disinformation countermeasures and artificial intelligence

Topic editors

Oleksandr Makarenko — Kyiv Polytechnic Institute, Ukraine
Paul Vines — Two Six Technologies, United States
George Cybenko — Dartmouth College, United States
Ludmilla Huntsman — Cognitive Security Alliance, United States

Citation

Makarenko, O., Vines, P., Cybenko, G., Huntsman, L., eds. (2025). *Disinformation countermeasures and artificial intelligence*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-7228-3

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 11 December 2025
DOI 10.3389/frai.2025.1750972

OPEN ACCESS

EDITED AND REVIEWED BY
Jize Zhang,
Hong Kong University of Science and Technology, Hong Kong SAR, China

*CORRESPONDENCE
Ludmilla H. Huntsman
ludmilla.huntsman@gmail.com

RECEIVED 20 November 2025
ACCEPTED 25 November 2025
PUBLISHED 11 December 2025

CITATION
Huntsman LH (2025) Editorial: Disinformation countermeasures and artificial intelligence. *Front. Artif. Intell.* 8:1750972. doi: 10.3389/frai.2025.1750972

Editorial: Disinformation countermeasures and artificial intelligence

Ludmilla H. Huntsman*

Cognitive Security Alliance, Alexandria, VA, United States

KEYWORDS

disinformation, information integrity, AI, cognitive security (CogSec), cognitive warfare, large language models (LLMs), cognitive resilience

Editorial on the Research Topic
Disinformation countermeasures and artificial intelligence

"Wars begin in the minds of men …," wrote U Thant, Secretary-General of the United Nations in 1968. The insight behind this statement—that while language structures reality, conflict takes shape first through narratives, ideas, and belief systems—remains no less relevant today. Humans have studied the relationship between thought, language, and mind for at least 2,500 years. In ancient times, Plato and Aristotle looked into how words relate to mental concepts and reasoning. During the Middle Ages and Early Modern period, Descartes, Locke, Leibniz, and Kant linked mental structure, representation, and logic, laying foundations for modern theories of knowledge, computation and cognition. Over the past century, this long-standing inquiry has taken shape in a diverse range of disciplines: philosophy of mind, cognitive psychology, neuroscience, psycholinguistics, artificial intelligence, and computational cognitive modeling, among others. With the rapid advancement of large language models and a race for artificial general intelligence, these fields have converged in the strategic domain of Cognitive Security (CogSec) to address the challenges of information integrity, cognitive warfare, and malign influence. State and non-state actors alike have weaponized linguistic framing, narrative engineering, and synthetic media generation in a global contest for epistemic authority: a war for reality. CogSec seeks to protect human information processing, belief formation, and decision-making by strengthening societies' cognitive resilience against disinformation, distorted reality and coercion carried out through information ecosystems.

Why does this research topic matter? Its significance emerges from a stark reality: the stakes of synthetic disinformation—systematically coordinated, AI-powered and amplified by bad actors—are not only epistemic or political. They are human, material, and often lethal. As I write this editorial, Russian soldiers launch missiles, drones, and guided bombs on Ukrainian cities for the fourth consecutive year. Russian state media justifies these war crimes domestically through narratives rooted in persistently distorted facts, heavily manipulated language and beliefs cultivated and reinforced by long-running state-directed disinformation campaigns. The tragedy illustrates an ugly truth: biased beliefs are algorithmically engineered and deployed at national scale can precipitate genocide and crimes against humanity. Disinformation kills, carries massive human suffering, and is an imminent threat to global security. It provokes and exacerbates conflict, erodes social cohesion, undermines trust in democratic institutions, and weakens societal resilience. The *Disinformation Countermeasures and AI* topic collection illustrates that while CogSec has become a critical domain, further research is needed to devise effective strategies on how to contain malign influence in the rapidly changing world.

When we began developing this Research Topic collection, the global information environment was already showing signs of destabilization. Yet during the span of its completion, the landscape has transformed more profoundly than anticipated. The acceleration of generative AI has altered not only the scale but the texture of disinformation, with interactive agents customizing and mimicking authenticity with increasing precision. Moreover, major geopolitical actors have escalated their use of information operations as instruments of statecraft. Meanwhile, the United States responded to this rapidly evolving threat with what experts described as unilateral disarmament and even surrender. After the closure of the U.S. government's main vehicle for countering foreign disinformation (GEC), along with the U.S. Agency for Global Media and other institutions, the global information sphere became even more vulnerable to malign influence operations and asymmetrically contested. With adversaries deliberately targeting cognitive, social, and institutional fault lines, this widening imbalance underscores why new research, new alliances, and new countermeasures are indispensable.

Our Research Topic will expand your understanding of the large, interdisciplinary spectrum of the topics within the field. Deepest thanks to my co-editors George Cybenko, Alexander Makarenko, and Paul Vines for their insight, leadership, and commitment to advancing this field. We extend our gratitude to all authors from Ukraine, Germany, France, United States, United Arab Emirates, United Kingdom, Bulgaria, Greece, Italy, and Switzerland who contributed to this research topic, to the reviewers whose expertise strengthened the scholarly quality of the collection, and to the editorial staff at *Frontiers in Artificial Intelligence*, *Frontiers in Big Data* and *Frontiers in Political Science: Politics of Technology* for their continuous support.

The ten peer-reviewed articles trace a coherent arc from conceptual foundations to concrete technical and policy responses to disinformation. Thompson and Guillory's history of the semantic hacking project distills lessons for modern cognitive security, while Deppe and Schaal's conceptual analysis of NATO's cognitive warfare framework clarifies the strategic terrain on which manipulation campaigns unfold. Paziuk et al. decode manipulative narratives in the Russia–Ukraine conflict and Zakharchenko shows how connective strategic narratives can bolster resilience, as Pilati and Venturini provide a worldwide mapping of how AI is already used in counter-disinformation practice. Romanishyn et al. and Marushchak et al. translate these insights into policy, offering recommendations for democratic resilience and regulatory lessons from Ukraine. At the technical edge, Dyachenko et al. explore LLM services for managing social communications, Tzoumanekas et al. propose a graph neural architecture search for bot detection, and Lipianina-Honcharenko et al. introduce OLTW-TEC, an online text-ensemble method for fake-news detection. Together, these contributions converge on a clear conclusion: effective counter-disinformation demands a whole-of-society approach, in which information integrity is achieved through advanced AI methods, attribution, public-private partnerships for cognitive resilience building, and adaptive democratic governance.

The challenge before us is not merely to develop more sophisticated classifiers or improved detection algorithms. It is to create cross-sector alliances to weave technology, education, societal values, and institutional frameworks into a trustworthy ecosystem. Researchers, practitioners, policymakers, and platform designers must work together to share best practices, develop transparent evaluation standards, and build open datasets and multimodal benchmarks. The work gathered in this Research Topic underscores the complexity of this challenge while pointing to pathways for technological, cognitive, and institutional innovation.

Our hope is that this Research Topic not only offers rigorous scholarship but also serves as a foundation for collective action and a catalyst for global collaboration. In a world where the integrity of information is continually tested, strengthening our cognitive and societal resilience is not merely an academic endeavor—it is a moral and strategic imperative.

Author contributions

LH: Writing – original draft, Writing – review & editing.

Conflict of interest

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

Generative AI statement

The author declares 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

frontiers | Frontiers in Artificial Intelligence

TYPE Methods
PUBLISHED 11 September 2024
DOI 10.3389/frai.2024.1401126

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Wayne Goodridge,
The University of the West Indies St. Augustine, Trinidad and Tobago

*CORRESPONDENCE
Khrystyna Lipianina-Honcharenko
✉ kh.lipianina@wunu.edu.ua

RECEIVED 14 March 2024
ACCEPTED 30 August 2024
PUBLISHED 11 September 2024

CITATION
Lipianina-Honcharenko K, Bodyanskiy Y, Kustra N and Ivasechko A (2024) OLTW-TEC: online learning with sliding windows for text classifier ensembles.
Front. Artif. Intell. 7:1401126.
doi: 10.3389/frai.2024.1401126

OLTW-TEC: online learning with sliding windows for text classifier ensembles

Khrystyna Lipianina-Honcharenko[1]*, Yevgeniy Bodyanskiy[2], Nataliia Kustra[3] and Andrii Ivasechko[1]

[1]Department of Information Computer Systems and Control, West Ukrainian National University, Ternopil, Ukraine, [2]Department of Artificial Intelligence, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine, [3]Department of Publishing Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine

In the digital age, rapid dissemination of information has elevated the challenge of distinguishing between authentic news and disinformation. This challenge is particularly acute in regions experiencing geopolitical tensions, where information plays a pivotal role in shaping public perception and policy. The prevalence of disinformation in the Ukrainian-language information space, intensified by the hybrid war with russia, necessitates the development of sophisticated tools for its detection and mitigation. Our study introduces the "Online Learning with Sliding Windows for Text Classifier Ensembles" (OLTW-TEC) method, designed to address this urgent need. This research aims to develop and validate an advanced machine learning method capable of dynamically adapting to evolving disinformation tactics. The focus is on creating a highly accurate, flexible, and efficient system for detecting disinformation in Ukrainian-language texts. The OLTW-TEC method leverages an ensemble of classifiers combined with a sliding window technique to continuously update the model with the most recent data, enhancing its adaptability and accuracy over time. A unique dataset comprising both authentic and fake news items was used to evaluate the method's performance. Advanced metrics, including precision, recall, and F1-score, facilitated a comprehensive analysis of its effectiveness. The OLTW-TEC method demonstrated exceptional performance, achieving a classification accuracy of 93%. The integration of the sliding window technique with a classifier ensemble significantly contributed to the system's ability to accurately identify disinformation, making it a robust tool in the ongoing battle against fake news in the Ukrainian context. The application of the OLTW-TEC method highlights its potential as a versatile and effective solution for disinformation detection. Its adaptability to the specifics of the Ukrainian language and the dynamic nature of information warfare offers valuable insights into the development of similar tools for other languages and regions. OLTW-TEC represents a significant advancement in the detection of disinformation within the Ukrainian-language information space. Its development and successful implementation underscore the importance of innovative machine learning techniques in combating fake news, paving the way for further research and application in the field of digital information integrity.

KEYWORDS

disinformation, fake news, online learning, classifier ensembles, machine learning

1 Introduction

During the russian-Ukrainian war, the importance of detecting fake news gains particular relevance, as disinformation can have serious consequences. The research by Tao and Peng (2023) and Mainych et al. (2023) emphasizes how social media, particularly Twitter and Weibo, play a key role in shaping public opinion during the conflict. The analysis shows that at the beginning of the war, there was a significant spike in activity, which later decreased but continued to draw attention to Ukraine as a victim of aggression. The importance of distinguishing be-tween real and fake news becomes even more critical, considering that social media is often used as the primary source of information. Thus, there is an increasing need for the development of effective methods for detecting and analyzing fake news to ensure the accuracy and safety of information. In this context, the development of effective methods for disinformation detection is critically important. This study offers an innovative approach that leverages the advantages of classifier ensembles combined with online learning methods. This approach allows not only for accurate identification of fake news but also for adaptation to the constantly changing information environment, ensuring the relevance and effectiveness of disinformation detection in real-time.

In addition to the growing relevance of disinformation detection during the russian-Ukrainian war, the broader research context highlights a gap in existing methodologies for addressing disinformation in non-English languages, particularly Ukrainian. While there has been substantial progress in the development of disinformation detection tools for English and other widely spoken languages, the unique linguistic and cultural characteristics of Ukrainian content require specialized approaches. This study seeks to fill this gap by developing a method that not only addresses the challenges of disinformation detection in the Ukrainian-language information space but also adapts to the dynamic and non-stationary nature of online news. By focusing on these specific needs, this research contributes to the global effort of combating disinformation and ensuring information integrity in regions severely affected by conflict. Furthermore, the study emphasizes the importance of a transparent data collection process, which is critical for the reproducibility of results and the broader applicability of the OLTW-TEC method in diverse linguistic contexts.

This research presents the development and analysis of the innovative method Online Learning with Sliding Windows for Text Classifier Ensembles (OLTW-TEC), which combines the advantages of adaptive online learning with a dynamic sliding window for text classifier ensembles. The goal is to increase the accuracy and adaptability in identifying fake news, particularly in the Ukrainian-language information space. This work highlights the importance of developing tools capable of adequately working with the unique linguistic and cultural features of specific language groups, as well as the necessity of ensuring the high responsiveness of systems in response to rapid changes in the information flow.

2 Related work

This comprehensive survey explores the landscape of fake news detection research, highlighting the diverse methodologies and technological innovations developed by scholars to tackle the ever-evolving challenge of identifying disinformation across various languages, platforms, and contexts.

Hamed et al. (2023) and Kondamudia et al. (2023) conduct review studies, analyzing various approaches to detecting fake news. Hamed and co-authors focus on challenges associated with datasets, feature representation, and data integration, while Kondamudia and co-authors examine attributes, features, and methods of detecting fake news in social networks, including linguistic and semantic analysis. In the studies by Hu et al. (2022) and Phan et al. (2023), the focus is on the use of deep learning and graph neural networks for detecting fake news. Hu and co-authors consider various deep learning approaches, including supervised, weakly supervised, and unsupervised learning, and analyze their effectiveness based on different datasets. Phan and co-authors focus on the use of graph neural networks, pointing out their potential and challenges related to data standardization and the development of specialized hardware.

The study by Das and Tsb (2023) highlights the importance of multi-contextual learning in the study of disinformation, considering various contexts such as content, emotions, users, and others. The authors offer a comprehensive view of this issue, including challenges associated with the multimodality of content and the scarcity of labeled data. Ruffo et al. (2023) also emphasize the necessity of an interdisciplinary approach, including network and language analysis, to understand the dynamics of the spread of fake news and its impact.

In the article by Baker et al. (2023), the way people express their feelings on Twitter during the conflict between Russia and Ukraine is investigated. The study has two objectives: to collect unique data and to use machine learning (ML) to classify tweets depending on their impact on people's emotions. The first goal was to identify the most relevant hashtags related to the conflict to find a dataset. The second goal was to use several well-known ML models to group tweets. Experimental results showed that most ML classifiers have higher accuracy on a balanced dataset. However, the results of experiments using data balancing strategies do not necessarily indicate that all classes will perform better. Therefore, it's important to emphasize the im-portance of comparing and contrasting data balancing strategies used in SA and ML research, including more classifiers and a broader range of use cases.

In the article, the authors Chang et al. (2024) introduce a new approach to detecting fake news using deep learning, employing natural language processing (NLP) techniques for encoding nodes with the context of news and users. They implement three graph convolutional networks to extract informative features from the news dissemination network and aggregate both internal and external user information. The methodology includes a global attention mechanism with memory for learning the structural homogeneity of news dissemination networks and a module for aggregating partial key information. Experimental results demonstrate the effectiveness of the proposed approach in detecting fake news, achieving high accuracy and F1-score metrics on real datasets. For instance, the GCN-GANM model showed high accuracy (Acc.: 0.9825, F1: 0.9825) on the Gossipcop dataset and (Acc.: 0.9804, F1: 0.9805) on the Weibo dataset, outperforming other methods. This work proposes a new direction in fake news detection research, combining global and partial information.

In the article, the authors Peng et al. (2024) developed a new method for detecting fake news that considers contextual semantic representation for analyzing multimodal data in social networks. This

method, named CSFND (Contextual Semantic representation learning for multimodal Fake News Detection), includes an unsupervised learning phase of context to obtain local contextual features of news, which are then combined with global semantic features for learning the news' contextual semantic representation. Experiments on two real multimodal datasets showed that CSFND significantly outperforms 10 state-of-the-art fake news detection methods, improving the average accuracy by 2.5% compared to the best baseline methods.

In the article by Ahammad (2024), the spread of fake news about COVID-19, which has become a critical issue, is investigated. The study aims to gain insights into the types of disinformation being spread and to develop a deep analytical approach for analyzing fake news about COVID-19. It combines sentiment analysis and topic modeling to improve the accuracy of extracting themes from a large volume of unstructured texts, considering the sentiment of words. A dataset containing 10,254 news headlines from various sources was collected and prepared, and rule-based SA was applied to tag the dataset with three sentiment tags. Among the evaluated TM models, Latent Dirichlet Allocation showed the highest coherence score of 0.66 for 20 clustered themes with negative sentiment and 0.573 for 18 clustered positive fake news themes, outperforming Non-negative Matrix Factorization (coherence: 0.43) and Latent Semantic Analysis (coherence: 0.40). The identified themes highlight that disinformation predominantly revolves around COVID vaccines, crimes, quarantine, medicine, and political and social aspects. This study provides insights into the impact of fake news about COVID-19, offers a valuable method for detecting and analyzing disinformation, and underscores the importance of understanding the patterns and themes of fake news to protect public health and promote scientific accuracy.

In the article by Qu et al. (2023), a new model for detecting fake news on social networks was developed, based on quantum multimodal fusion (QMFND). The QMFND model integrates extracted image and text features, which go through a proposed quantum convolutional neural network (QCNN) to obtain discriminative results. Testing QMFND on two social media datasets, Gossip and Politifact, showed that its performance is equal to or even exceeds that of classic models. Furthermore, the impact of various parameters was explored. QCNN proved to be not only highly expressive and capable of entanglement but also resistant to quantum noise.

In another article, Farhangian et al. (2024) conducted a detailed study in the field of fake news detection, proposing an updated taxonomy for this domain based on several criteria: types of used features, perspectives on fake news detection, methods of feature representation, and approaches to classification. The authors conducted a wide-ranging empirical study, evaluating various feature representation techniques and classification approaches based on accuracy and computational costs. Experimental results showed that optimal feature extraction techniques depend on the characteristics of the dataset. Transformer-based models consistently demonstrated higher performance. Moreover, using transformer models as feature extraction methods, rather than merely fine-tuning the network for post-activity, improves overall performance. Through careful error analysis, it was discovered that a combination of feature representation methods and classification algorithms, including classical ones, offer complementary aspects and should be considered to achieve better overall performance while maintaining relatively low computational costs.

In the article by Fang et al. (2024), a new approach to the early detection of fake news through the perception of the news semantic environment (NSEP) is introduced. NSEP utilizes graph convolutional networks to detect semantic inconsistencies between the content of news and external posts, as well as a microsemantic detection module with multi-head and sparse attention to identify semantic contradictions. Experiments on real Chinese and English datasets showed that NSEP achieves an accuracy of up to 86.8% on Chinese datasets, which is 14.1% higher than other methods. This confirms the effectiveness of detecting fake news through the analysis of micro- and macrosemantic environments.

In the article by Soga et al. (2023), the problem of detecting fake news on social networks is examined. The authors propose a new approach that considers the similarity of user opinions by analyzing their positions regarding news articles and interactions in posts. Using a network of graph transformers, the method simultaneously extracts global structural information and interactions of similar positions. The method was evaluated on specially collected data from Twitter and the FibVID dataset, demonstrating significant improvement compared to traditional methods, including state-of-the-art approaches.

In the article by Yang et al. (2024), the problem of detecting fake news in multimodal data is examined. The authors highlight the challenges related to analyzing the relationships between regions of images and fragments of text, as well as the necessity for a deeper analysis of hierarchical text semantics. To address these issues, a Multimodal Relation Attention Network (MRAN) is proposed, which includes several processing stages. Initially, a multi-level encoding network is used to extract semantic features of the text, along with VGG19 for extracting visual characteristics. Then, an attention network is utilized to compute the similarity between segments of information within and across modalities. Finally, the obtained features are fed into a fake news detector. Experiments on three datasets demonstrated the effectiveness of MRAN, underscoring its strong performance in detecting fake news.

In the article by Raja et al. (2024), the problem of detecting fake news in Dravidian languages, which are low-resourced, is discussed. The authors propose a hybrid deep learning model that integrates Enhanced Temporal Convolutional Neural Networks (DTCN), Bidirectional Long Short-Term Memory (BiLSTM), and Contextualized Attention Mechanism (CAM). DTCN is used for capturing temporal dependencies, BiLSTM for effectively capturing long-term dependencies, and CAM for emphasizing important information and reducing the impact of irrelevant content. The model also employs an adaptive cyclic learning rate with an early stopping mechanism to improve model convergence. The results show that the proposed model outperforms existing methods and achieves a high average accuracy of 93.97% on the Dravidian_Fake dataset in four Dravidian languages.

In the article by Luvembe et al. (2024), the issue of detecting fake news in a multimodal context is explored. The authors point out the shortcomings of existing methods that integrate cross-modal features without considering uncorrelated semantic representations, which can introduce noise into the multimodal characteristics. This lowers the accuracy of models, as it is crucial to account for the subtle differences between text and images in identifying fake news. To overcome these

challenges, the CAF-ODNN model is proposed, which utilizes complementary attention and an optimized deep neural network to detect subtle cross-modal connections. The model includes generating captions for images for semantic representation, bidirectional complementary attention between modalities, and an alignment and normalization component for calibrating fused representations. The Optimized Deep Neural Network (ODNN) is used to enhance feature extraction. The model outperforms similar approaches on standard metrics across four real-world datasets, highlighting the importance of complementary attention and optimization in detecting fake news.

The article by Jiang et al. (2023) examines a new approach to detecting fake news, which employs multimodal learning using query prompts. Traditionally, fake news detection was performed through the analysis of textual information, but this approach often overlooks the complexity and nuances of online disinformation. The new method proposed in the article includes the use of multimodal data (text and images) and a training methodology using query prompts, which allows for better detection of fake news, especially in situations with limited data. The authors utilize three types of query prompts with a soft verbalizer and a merging method that considers similarity for adaptively combining multimodal representations. This approach shows higher F1 scores and accuracy on two multimodal standard datasets, confirming its effectiveness in real-world scenarios.

The article by Syed et al. (2023) proposes a hybrid approach that combines weakly supervised learning and deep learning for detecting fake news on social networks. The study focuses on using machine learning methods, such as SVM, for annotating large volumes of unlabeled data, as well as applying deep learning techniques like Bi-LSTM and Bi-GRU for classifying fake news. The authors utilize TF-IDF and Count Vectorizer for feature extraction from textual data. Experimental results show that the proposed approach achieves high accuracy in detecting fake news, making it an effective tool in combating online disinformation.

The article by Xie and Li (2023) introduces the concept of "gatekeepers" in social networks for detecting fake news. Gatekeepers are active users who participate in the dissemination of news. The research proposes a gatekeeper behavior model based on Recurrent Neural Network (RNN), which includes training the model and detecting fake news. The method is capable of detecting fake news in real-time using data from Twitter and Weibo. Experimental results demonstrate that the Gated Recurrent Unit (GRU) achieves the best overall performance. The proposed method surpasses several contemporary approaches, showcasing its effectiveness in the early and middle stages of news dissemination.

In the article by Přibáň et al. (2019), the task of auto-mating the detection of fake news and fact-checking for West Slavic languages, specifically Czech, Polish, and Slovak, is addressed. The authors present datasets for these languages and conduct preliminary experiments that establish a baseline for further research in this area. They utilize 10-fold cross-validation to evaluate both balanced and unbalanced datasets, as well as conduct binary experiments with just the "TRUE" and "FALSE" classes. The input data for the classifier consists of either the statement text alone or the statement text supplemented by justification text.

In the article by Bucos and Drăgulescu (2023), the efficiency of using the back-translation (BT) method with transformer models to improve the detection of fake news in Romanian is explored. The study is based on data from Factual.ro, where models with BT showed better results in terms of accuracy, precision, recall, F1-score, and AUC compared to models trained on the original dataset. The use of mBART for BT with French as the target language improved the model's performance compared to Google Translate. The Extra Trees Classifier and the Random Forest Classifier were among the best-performing models tested. The results indicate the potential of using BT with transformer models like mBART to enhance the effectiveness of fake news detection.

In the article by Afanasieva et al. (2022), the problem of determining the veracity of information, especially in the context of social unrest and significant events such as the US presidential elections and Russia's invasion of Ukraine, is investigated. The authors examine the efficiency of using neural networks to detect fake news. To improve classification accuracy, a data preprocessing algorithm based on the fundamental principles of natural language processing was developed. The study identified linguistic patterns of fake news, which became the basis for data preprocessing. The features of convolutional and recurrent neural networks and their modifications for analyzing textual data are described. A set of metrics was chosen to compare certain models, characterizing the efficiency of algorithms. The accuracy of these models was tested on data related to the US presidential elections and the large-scale invasion of the russian into Ukraine.

Supplementary Table 1 presents a comparative analysis of different studies in the field of fake news detection using various datasets, models, languages of datasets, and approaches.

Based on the analysis of contemporary research in the field of disinformation detection presented in Supplementary Table 1, it can be concluded that existing approaches significantly focus on the English and Chinese languages, leaving the Ukrainian language with an insufficient level of attention. This indicates a lack of specialized datasets and models that would be adapted to the peculiarities of the Ukrainian language. In this context, the approach "Online Learning with Sliding Windows for Text Classifier Ensembles" (OLTW-TEC) has been developed, aimed at creating a comprehensive method for detecting disinformation in Ukrainian-language content. The method includes stages of data collection and preprocessing, analysis of sentiment, emotions, and text vectorization, allowing for a deeper analysis and more effective detection of fake news, relying on the unique linguistic and cultural features of the Ukrainian language.

3 Meta-model—OLTW-TEC

Let us introduce the enhancement of an ensemble authors Bodyanskiy et al. (2024) of adaptive predictors for multidimensional non-stationary sequences and its online training, aimed at improving efficiency in the context of text classification. This is achieved by integrating advanced natural language processing methods and adaptive learning algorithms. Specifically, vector characteristics of text documents, optimally selected for the given task, are used as input data for the classification model ensemble. The optimization of these models' weights in the ensemble is performed using the Adam algorithm. An important aspect of the enhancement is the implementation of the "sliding window" method, which ensures the adaptability of the predictors to changes in the data. Such an approach allows for high accuracy and adaptability of models under the dynamic conditions of online learning.

Let us consider an ensemble of models for text classification $MP_1, \supset, MP_j, \supset, MP_h$, each processing the vector characteristics of

a text document obtained using the best model of vectorization. These characteristics are represented as $x(\tau)=(x_1(\tau),\ldots,x_i(\tau))^T$ for $\tau=1,2,\ldots,T$. The estimate that appears at the output of each member of the ensemble will be denoted as $x(\hat{\tau})_j$ for $h_j=1,2,\ldots,h$, where each $x(\hat{\tau})_j$ represents the class score for the corresponding document. The members of the ensemble can include both traditional text classification models such as logistic regression or SVM, as well as more complex neural networks, including recurrent networks and deep learning structures like LSTM or transformers.

The estimates $x(\tau)_j$ from each model of the ensemble are fed into the meta-model, which forms a combined prediction for text classification. The prediction of the meta-model is presented in the form

$$x^*(\tau)=\sum_{j=1}^{h}c_j x(\hat{\tau})_j = x(\hat{\tau})c,$$

where $c=(c_1,\ldots,c_j,\ldots,c_h)^T$ is a weight vector that determines the contribution of each individual model in the ensemble.

The matrix $x(\hat{\tau})=\left(x(\hat{\tau})_1,\ldots,x(\hat{\tau})_j,\ldots,x(\hat{\tau})_h\right)$ is formed from the class estimates generated by each model. The parameters of the meta-model satisfy the condition of unbiasedness:

$$\sum_{j=1}^{h}c_j = c^T E_h = 1,$$

where E_h is a vector formed by ones. The weights c are adjusted in such a way as to optimize the overall accuracy of the meta-model's classification.

To determine the optimal parameters of the meta-model (weight vector c), the Adam optimization algorithm is used, an alternative to traditional methods such as the Lagrange multipliers. The loss function $L(c)$, which is minimized, is defined as the sum of the squares of the difference between the true classes and the predictions of the meta-model:

$$L(c)=\sum_{\tau=1}^{T}\| x(\tau)-x(\hat{\tau})c\|^2,$$

where $x(\tau)$ represents the true class labels. Minimizing the loss function $L(c)$ using Adam allows for the efficient adjustment of the weights c, ensuring the optimal combination of predictions from different ensemble models to achieve high classification accuracy.

In cases of time-varying or non-stationary textual data, the efficiency of the meta-model can be enhanced using the "sliding window" method. This method involves updating the parameters of the meta-model using only the most recent s observations (text documents) from $x(T-s+1)to\,x(T)$. When a new observation $x(T+1)$ arrives, the oldest observation in the "window" is removed, and the assessment is based on data from $x(T-s+2)to\,x(T+1)$ This approach allows the meta-model to be more adaptable to changes in data patterns and improves its predictive ability based on current information, especially in situations where textual data is characterized by high dynamics or non-stationarity.

Choosing the optimal size of the "sliding window" s is crucial for achieving the highest classification accuracy in the ensemble of meta-models. This choice often relies on empirical considerations, as *a priori* knowledge about the nature of changes in textual data can be limited. In scenarios where different "sliding window" sizes may be optimal for different types of textual data, an effective approach is to create a set of meta-models, each built for a specific window size.

To determine the best meta-model, a second-level meta-model can be applied, which evaluates and selects the most efficient meta-model based on its performance across the entire training sample. This approach allows for dynamic adaptation to changes in textual data and selecting the best classification method depending on the specific context.

4 Method

The developed comprehensive method for detecting disinformation covers everything from data collection and preprocessing to sentiment analysis, emotion analysis, and text vectorization. This method includes the application of advanced machine learning techniques and neural networks, providing a deep analysis of textual data and effective detection of fake news. Supplementary Figure 1 illustrates the structure of the proposed method as a series of the following steps:

4.1 Step 1: data collection (block 1)

Data collection is a critically important stage in the process of developing a method for detecting disinformation. Properly collected and structured data allow for the effective training of machine learning models and analytics. Here is a more detailed description of this step:

4.1.1 Collection of textual data

1 Source identification: determining sources for data collection is an important task. Sources can include news portals, social networks, blogs, forums, and other platforms where users can publish or share information.
2 Data collection automation (Gramyak et al., 2022): developing scripts or using existing tools for automatic data collection. This can include web scraping, social network API requests, and more.
3 Data filtering and validation: filtering and validating data to ensure their quality and relevance to research requirements.

4.1.2 Metadata collection

1 Author information: collecting data about the authors of texts may include information about their profiles, publication history, number of followers, and other social indicators that can be useful for analysis.
2 Source information:. collecting information about the sources of texts, such as URL, publication date, number of views, likes, comments, and other indicators that can be important for analyzing the context and popularity of publications.
3 Structuring metadata. organizing collected metadata into structured databases for easy access and analysis in later stages of research.

4.2 Step 2: data preprocessing (block 2)

Data preprocessing is a fundamental step in the process of analyzing textual information. This step involves various techniques and methods that help prepare the collected data for further analysis. Here is a more de-tailed description of this step (Lipianina-Honcharenko et al., 2022a; Gramyak et al., 2022):

1 Tokenization (Lipianina-Honcharenko et al., 2023): the process of breaking text into individual words, phrases, symbols, or other meaningful elements called tokens. This assists in further analysis and processing of the text.
2 Stemming (Lipyanina et al., 2020): the process of removing suffixes, prefixes, and infixes from words to return them to their base form. This facilitates the detection of common themes and patterns in the text.
3 Part-of-speech tagging: the process of identifying the parts of speech of each word in the text, which can be useful for syntactic analysis and determining semantic relationships between words.
4 Named entity recognition: identifying and classifying named entities in the text, such as names of people, organizations, locations, etc.

4.3 Step 3: text vectorization (block 3)

Text vectorization is a crucial step that trans-forms textual data into a numerical format, convenient for analysis and processing using machine learning methods (Lipyanina et al., 2020). Let us look more closely at this process:

4.3.1 Choosing the vectorization method

1 Word2Vec (Golovko et al., 2019): this model trains vector representations of words in a multidimensional space such that words that frequently occur together have similar vector representations.
2 GloVe:. another approach to word vectorization that uses both local and global statistical analysis of the text corpus to determine vector representations of words.
3 BERT: a modern model that uses attention mechanisms to determine relationships between words in text and can learn deep contextual representations of words.

4.3.2 The vectorization process

1 Training or loading models: models can be trained on data or use pre-trained models for text vectorization.
2 Transforming text: applying the chosen model to transform each word in the text into vector representations.

4.3.3 Building vector representations

1 Word vectorization: obtaining vector representations for each individual word in the text.
2 Text vectorization: aggregating vector representations of words to obtain vector representations of entire texts. This can be done through averaging, summing, or other aggregation methods.

4.4 Step 4: sentiment and emotion analysis (block 4)

Sentiment and emotion analysis is key to understanding the mood and nuances contained in textual data. This can help determine whether a text is positive, negative, or neutral, as well as identify potential emotional responses that may be associated with disinformation. Here's a more detailed description of this step:

4.4.1 Choosing models for sentiment and emotion analysis

Sentiment and emotion analysis, as outlined by Lipianina-Honcharenko et al. (2022), is crucial for understanding the mood and subtleties in textual data, particularly in identifying whether a text is positive, negative, or neutral, and recognizing emotional responses linked to disinformation.

1 Ready-made models: there are many pre-trained models for sentiment and emotion analysis, such as vader, textblob, or models based on bert and other deep neural networks.
2 Custom models: depending on the specific case, it may be useful to develop and train custom models on data specific to the task.

4.4.2 Sentiment analysis

1 Calculating sentiment: applying models to determine the sentiment of each text, allowing to determine whether a text is positive, negative, or neutral.
2 Interpreting results: analyzing results to determine the overall mood of the data and identify possible anomalies or patterns.

4.4.3 Emotion analysis

1 Calculating emotional tone: applying models to determine the emotional tone of the text, such as joy, sadness, anger, surprise, fear, etc.
2 Interpreting results: analyzing the obtained emotional tones to determine how emotions may be associated with disinformation and how they can be used for further analysis.

4.5 Step 5: online learning with sliding window for text classifier ensembles (block 5)

This step focuses on developing and training a classification model capable of detecting fake information based on text analysis and other identified features. Here's a detailed description of this step:

4.5.1 Creation and training of model ensemble

1 Creating an ensemble of models: selecting and creating various classification models (e.g., logistic regression, svm, lstm, transformers).
2 Training models: each model is trained separately on vectorized textual data.

3 Forming a meta-model: using an algorithm, such as adam, to optimize the weights of models in the ensemble, ensuring better integration and selection of predictions from each model.

4.5.2 Implementing the "Sliding Window" method for online learning

1 Implementing "Sliding Window": setting the size of the "sliding window" to select the most recent text documents that will be used for continuous updating and training of the models.
2 Online updating of models: periodically updating the ensemble models using the latest incoming data and discarding outdated data to maintain relevance and high prediction accuracy.
3 Adapting to changes in data: continuously adapting the meta-model to changes in textual data, ensuring an effective response to the dynamism and non-stationarity of text sequences.

4.6 Step 6: adaptation and retraining (block 8)

This stage aims to ensure the system's ability to adapt to the evolution and change in forms of disinformation, guaranteeing its prolonged effectiveness in combating fake news (Golovko et al., 2019). Let us consider the details of this step:

4.6.1 Retraining the system

1 Collecting new data: continuously gathering new data from open sources to reflect the latest trends and patterns of disinformation.
2 Assessing the need for retraining: analyzing the current efficiency of the system and determining if there is a need for retraining based on new data.
3 Retraining the model: applying the training process to the model using new data to adapt the model to new forms of disinformation.

4.6.2 Updating models and algorithms

1 Analyzing new algorithms and technologies: evaluating and analyzing new algorithms and technologies that could be used to improve system efficiency.
2 Updating algorithms: making changes to the algorithms and methods used based on the information obtained and analysis of results.
3 Testing and validating updated models: conducting testing and validation of updated models to ensure their effectiveness and reliability.

4.6.3 Monitoring and evaluation

1 Monitoring system efficiency: constantly monitoring the efficiency of the system to identify potential problems or areas for improvement.
2 Feedback and adaptation: collecting and analyzing feedback from users and experts for further improvement and adaptation of the system.

The comprehensive method for detecting disinformation outlined effectively integrates advanced machine learning techniques and neural networks, offering a highly adaptable and accurate tool for combating fake news through meticulous data collection, preprocessing, sentiment and emotion analysis, text vectorization, and the innovative application of online learning with sliding window techniques for text classifier ensembles.

5 Results

For the implementation of the proposed method, a dataset available on Kaggle (Ukrainian news, 2024) was chosen, which is a unique collection of approximately 60,000 news headlines collected from February 24 to December 11, 2022, covering the period of the full-scale russian-Ukrainian war. It includes both verified and fake news, collected from Ukrainian Telegram channels and russian channels with fakes, making it the largest open source of relevant data. The dataset contains two main attributes: the text of the news headline and a label indicating the veracity (True for confirmed news and False for fake news). In total, it contains 4,522 records with the label 'False'. The data sources included Telegram channels such as "SUSPILNE NEWS," "Perepichka NEWS," and others.

To ensure the accuracy and completeness of the collected data, the process of data collection is carefully structured. The dataset was curated using automated scripts for data scraping and API requests to gather news headlines from Ukrainian and Russian Telegram channels, which are known to be primary sources of both verified and fake news during the russian-Ukrainian war. Specific criteria were established to select reliable sources, such as official news outlets like "SUSPILNE NEWS" and well-known channels that have been identified as disseminators of disinformation, like "Perepichka NEWS." The selection process also involved filtering out irrelevant content to maintain the dataset's focus on news items directly related to the conflict.

In the implementation phase, data was initially prepared, where textual materials were converted into a numerical format using the TF-IDF vectorization method. Subsequently, several individual models were trained based on these data, including logistic regression, SVM, random forest, gradient boosting, KNN, decision tree, XGBoost, and AdaBoost. Each model was adapted and optimized to solve the classification task using the training dataset. In the second phase of implementation, a meta-model based on XGBoost was formed, which was trained using predictions obtained from individual models through a stacking mechanism. The classification accuracy was evaluated on a test dataset, with results presented through a classification report and confusion matrix, visualized as a heatmap. This approach demonstrates the high potential of ensemble methods and stacking to enhance machine learning efficiency in complex text classification tasks, particularly in detecting disinformation.

The obtained classification results are presented in a report format (Supplementary Figure 2), which includes metrics such as precision, recall, f1-score, and support for two classes: "False" (fake news) and "True" (true news).

In the process of evaluating the effectiveness of the news classification model (Supplementary Figure 2A), an analysis was conducted on its ability to differentiate factual truths from fake

messages. According to the results, the model shows a high overall classification accuracy of 93%. For the "False" class, which is presumed to identify fake news, the precision was 0.95, indicating a high probability of correctly classifying a news item as fake when the model does so. Meanwhile, the recall for this class is 0.72, indicating that about 28% of fake news was missed by the model. On the other hand, for the "True" class, corresponding to true news, the model showed an impressive recall level of 0.99, meaning it successfully identified 99% of true news. Such an ability of the model to effectively classify true news is an important feature in the context of combating disinformation.

The analytical approach to evaluating the model's performance also included an analysis of the confusion matrix. In this matrix (Supplementary Figure 2B), 1,793 true negative results were noted, confirming the correct classification of fake news. However, 705 false negative results were identified, indicating the existence of a certain number of fake news that the model mistakenly classified as true. For the true news class, 8,138 true positive results and 99 false positives were registered, demonstrating high accuracy in detecting factually true content. The F1-score for both classes, which harmonizes precision and recall, was found to be 0.82 for "False" and 0.95 for "True," overall confirming the model's balance and reliability.

Analyzing the results presented in Supplementary Table 2, scientifically substantiated conclusions can be made about the effectiveness of the OLTW-TEC method compared to classic classification approaches. Examining each metric individually, it is evident that the OLTW-TEC method demonstrates significant improvements across most parameters.

In the context of precision for the "False" class, OLTW-TEC achieves a value of 0.95, which is comparable to logistic regression and surpasses other classic methods, except for Gradient Boosting. This indicates OLTW-TEC's high ability to correctly identify fake news. Regarding the "True" class, OLTW-TEC's precision is 0.92, a competitive figure, especially compared to SVM and XGBoost.

The recall rate for the "False" class in OLTW-TEC at 0.72 is significantly higher than that of logistic regression and matches the performance of SVM, indicating the model's improved ability to detect fake news among the negative class. For the "True" class, the recall rate is 0.99, a common trend among all considered models, highlighting their capability to identify true news.

The harmonic mean between precision and recall, defined as the F1-score, for OLTW-TEC is 0.82 for the "False" class and 0.95 for the "True" class, effectively confirming the model's balance between these two crucial metrics. This significantly exceeds the F1-scores of logistic regression for the "False" class and is on par with the best results among other classic methods.

The overall accuracy of OLTW-TEC is 93%, which is one of the highest figures among all the models considered (Supplementary Table 2), confirming its strong position as a reliable tool for news classification. The confusion matrix also indicates a high number of correctly classified instances for both classes.

6 Discussion

Analyzing the results presented in Supplementary Table 1, significant achievements in the field of fake news detection using various approaches, ranging from deep learning to multi-modal analysis, were identified. However, an important gap in existing research was discovered, namely the insufficient attention to Ukrainian-language content. Our OLTW-TEC method, focused on a Ukrainian dataset, fills this gap, highlighting the importance of developing specialized solutions for specific linguistic and cultural contexts.

Comparative analysis with other studies (see Supplementary Table 1) showed that OLTW-TEC achieves significant results in accuracy, demonstrating values of 0.95 for the "False" class and 0.92 for the "True" class. These metrics indicate the method's high efficiency in detecting fake news and true messages, respectively. Particularly valuable is OLTW-TEC's ability to provide a high recall (recall) for the "False" class at 0.72, higher compared to some other methods, such as logistic regression. This indicates the method's effectiveness in the context of detecting fake news, which often have complex patterns and can be disguised as credible.

F1-scores, harmonizing precision and recall, also highlight OLTW-TEC among other studies, underscoring its balance and reliability. An overall accuracy of 93% demonstrates that OLTW-TEC can serve as a reliable tool in the fight against disinformation and in detecting fake news. The confusion matrix confirms the model's high accuracy with a minimal number of false classifications, which is especially important in situations where ensuring information accuracy is crucial to prevent panic or disinformation during critical moments.

It's worth noting that although OLTW-TEC shows significant results compared to classic methods, there are certain limitations that should be considered when interpreting these results. Firstly, the use of ensemble methods and the "sliding window" technique requires substantial computational resources, which may limit the widespread application of OLTW-TEC in real-time, especially on devices with limited computing power. This raises concerns about the scalability of the method, particularly when applied to large-scale datasets or deployed in environments with constrained computational resources.

Secondly, the method may be sensitive to the size of the "sliding window," as an incorrect choice of size can lead to overfitting or underfitting of the model, especially in conditions where data dynamics change rapidly. This requires further research to optimize parameters and improve the model's adaptability, ensuring that the method can maintain high accuracy across different contexts and data flows.

Thirdly, although OLTW-TEC demonstrated high efficiency on Ukrainian-language data, its universality and efficiency in datasets of other languages and cultural contexts have not been fully explored. Further experiments are needed to determine whether the method can maintain similar performance metrics in other linguistic environments. This would involve testing the model on multilingual datasets and assessing its robustness across diverse linguistic and cultural contexts.

Finally, considering the rapid development of technology and the changing nature of disinformation dissemination, there is a need for continuous updating and adaptation of the model to maintain its relevance. This concerns not only algorithmic improvements but also the collection and integration of new data for training, which in turn requires additional efforts to ensure the reliability and quality of the input data.

Given these considerations, it would be beneficial to address the scalability and computational efficiency of OLTW-TEC in future research. This could involve exploring methods to reduce computational load, such as model pruning, optimizing algorithmic complexity, or employing more efficient hardware solutions. Additionally, incorporating distributed computing techniques or leveraging cloud-based platforms could help manage the computational demands of the method, making it more accessible for broader applications.

Overall, while OLTW-TEC shows promising results, it is important to emphasize the need for continued development and adaptation of the method to effectively respond to the constantly changing challenges in the field of disinformation detection. Future work should focus on addressing the identified limitations to enhance the scalability, efficiency, and applicability of the proposed approach across different contexts and environments.

Future research will focus on enhancing the OLTW-TEC method, particularly on expanding its computational efficiency and adaptability. One of the priorities is the optimization of algorithms to reduce computational power requirements, allowing the method to be applied in a wider range of applications, including mobile and other de-vices with limited resources. Attention will also be focused on more precise tuning of the "sliding window" parameters to improve prediction accuracy in various conditions of dynamic data flow. Another important direction is the expansion of method valida-tion on diverse linguistic datasets, which will determine its universality and effective-ness in a global context. Additionally, mechanisms for continuous updating of the training dataset are planned to be implemented, allowing the system to adapt to new patterns and forms of disinformation. These measures aim not only to improve the accuracy and reliability of OLTW-TEC but also to ensure its resilience to rapid changes in the digital information space.

7 Conclusion

In the course of the conducted research, the effectiveness of the OLTW-TEC method was thoroughly examined and evaluated. This innovative approach addresses the challenge of text classification in the context of non-stationary data and dynamic online learning. The methodology incorporates the use of an ensemble of adaptive classifiers, vector representations of texts, optimization of classifier weights using the Adam algorithm, and the implementation of a "sliding window" mechanism to maintain the model's relevance.

According to the evaluation results presented in the classification report, OLTW-TEC demonstrated high accuracy, achieving a 93% performance rate. Precision and recall for the "False" class were found to be 0.95 and 0.72, respectively, indicating high efficiency in detecting fake news with a minimal percentage of false negatives. For the "True" class, the model provided almost perfect recall with a rate of 0.99, thereby confirming its ability to recognize true news. The obtained results illustrate that OLTW-TEC possesses a high balance between different metrics, ensuring reliable and effective classification in real-world conditions.

The analysis of the confusion matrix points to a high degree of classification accuracy, with a minimal number of false positives and false negatives. This demonstrates the reliability of OLTW-TEC as a tool for information filtering, which is particularly relevant in the context of combating disinformation.

Compared to traditional classification methods (see Supplementary Table 1), OLTW-TEC not only shows better results across most metrics but also provides room for adaptation to changes in the nature of the data. The choice of "sliding window" size and the possibility of adjusting it depending on the data specifics give the method additional flexibility and precision.

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

KL-H: Conceptualization, Data curation, Methodology, Project administration, Supervision, Writing – original draft. YB: Writing – review & editing, Conceptualization, Methodology. NK: Formal analysis, Validation, Writing – review & editing. AI: Data curation, Software, Visualization, Writing – original draft.

Funding

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

Acknowledgments

GPT4 was used to translate the text.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1401126/full#supplementary-material

References

Afanasieva, I., Golian, N., Golian, V., Khovrat, A., and Onyshchenko, K., (2022). "Application of neural networks to identify of fake news." in *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems. Volume II: Computational Linguistics Workshop Kharkiv, Ukraine*. 3396, 346–358. Available from: https://ceur-ws.org/Vol-3396/paper28.pdf (Accessed March 14, 2024).

Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: an in-depth sentiment analysis and topic modeling approach. *Nat. Lang. Proces. J.* 6:100053. doi: 10.1016/j.nlp.2024.100053

Baker, M., Jihad, K., and Taher, Y. (2023). Prediction of people sentiments on twitter using machine learning classifiers during russian aggression in Ukraine. *Jord. J. Comp. Inform. Technol.* 1:1. doi: 10.5455/jjcit.71-1676205770

Bodyanskiy, Y. V., Lipianina-Honcharenko, K. V., and Sachenko, A. O. (2024). Ensemble of adaptive predictors for multivariate nonstationary sequences and its online learning. *Radio Electron. Comp. Sci. Control.* 4:91. doi: 10.15588/1607-3274-2023-4-9

Bucos, M., and Drăgulescu, B. (2023). Enhancing fake news detection in romanian using transformer-based back translation augmentation. *Appl. Sci.* 13:13207. doi: 10.3390/app132413207

Chang, Q., Li, X., and Duan, Z. (2024). Graph global attention network with memory: a deep learning approach for fake news detection. *Neural Netw.* 172:106115. doi: 10.1016/j.neunet.2024.106115

Das, B., and Tsb, S. (2023). Multi-contextual learning in disinformation research: a review of challenges, approaches, and opportunities. *Online Soc. Networks Media* 34-35:100247. doi: 10.1016/j.osnem.2023.100247

Fang, X., Wu, H., Jing, J., Meng, Y., Yu, B., Yu, H., et al. (2024). NSEP: early fake news detection via news semantic environment perception. *Inf. Process. Manag.* 61:103594. doi: 10.1016/j.ipm.2023.103594

Farhangian, F., Cruz, R. M. O., and Cavalcanti, G. D. C. (2024). Fake news detection: taxonomy and comparative study. *Inform. Fusion* 103:102140. doi: 10.1016/j.inffus.2023.102140

Golovko, V., Kroshchanka, A., Komar, M., and Sachenko, A. (2019). "Neural network approach for semantic coding of words," in *Lecture notes in computational intelligence and decision making. ISDMCI 2019. Advances in intelligent systems and computing [online]*. eds. V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, S. Radetskaya (Cham: Springer International Publishing), 647–658.

Gramyak, R., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., and Zahorodnia, D., (2022). "Intelligent method of a competitive product choosing based on the emotional feedbacks coloring." in Proceedings of the 2nd international workshop on intelligent information technologies & systems of information security with CEUR-WS. 2853, 346–357. Available at: https://ceur-ws.org/Vol-2853/paper31.pdf (Accessed March 14, 2024).

Hamed, S. K., Ab Aziz, M. J., and Yaakub, M. R. (2023). A review of fake news detection approaches: a critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon* 9:e20382. doi: 10.1016/j.heliyon.2023.e20382

Hu, L., Wei, S., Zhao, Z., and Wu, B. (2022). Deep learning for fake news detection: a comprehensive survey. *AI Open* 3, 133–155. doi: 10.1016/j.aiopen.2022.09.001

Jiang, Y., Yu, X., Wang, Y., Xu, X., Song, X., and Maynard, D. (2023). Similarity-aware multimodal prompt learning for fake news detection. *SSRN Electron. J.* 201. doi: 10.2139/ssrn.4347542

Kondamudia, M. R., Sahoob, S. R., Chouhanc, L., and Yadavd, N. (2023). A comprehensive survey of fake news in social networks: attributes, features, and detection approaches. *J. King Saud Univ. Comp. Inform. Sci.* 35:101571. doi: 10.1016/j.jksuci.2023.101571

Lipianina-Honcharenko, K., Lendiuk, T., Sachenko, A., Osolinskyi, O., Zahorodnia, D., and Komar, M. (2022). "An intelligent method for forming the advertising content of higher education institutions based on semantic analysis," in *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science*. ed. O. Ignatenko (Cham: Springer International Publishing). 169–182.

Lipianina-Honcharenko, K., Savchyshyn, R., Sachenko, A., Chaban, A., Kit, I., and Lendiuk, T. (2022a). Concept of the intelligent guide with AR support. *Int. J. Comp.* 21, 271–277. doi: 10.47839/ijc.21.2.2596

Lipianina-Honcharenko, K., Wolff, C., Sachenko, A., Desyatnyuk, O., Sachenko, S., and Kit, I. (2023). Intelligent information system for product promotion in internet market. *Appl. Sci.* 13:9585. doi: 10.3390/app13179585

Lipyanina, H., Sachenko, O., Lendyuk, T., Sachenko, A., and Vasylkiv, N. (2020). "Intelligent method of forming the HR management short-term project," in *Advances in Intelligent Systems and Computing V. CSIT 2020*. eds. N. Shakhovska, M. O. Medykovskyy N. Shakhovska, M. O. Medykovskyy (Cham: Springer International Publishing), 1045–1055.

Luvembe, A. M., Li, W., Li, S., Liu, F., and Wu, X. (2024). CAF-ODNN: complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Inf. Process. Manag.* 61:103653. doi: 10.1016/j.ipm.2024.103653

Mainych, S., Bulhakova, A., and Vysotska, V., (2023). "Cluster analysis of discussions change dynamics on twitter about war in Ukraine." in Proceedings of the 7th international conference on computational linguistics and intelligent systems. Volume II: Computational linguistics workshop Kharkiv, Ukraine. 3396, 490–530. Available at: https://ceur-ws.org/Vol-3396/paper39.pdf (Accessed March 14, 2024).

Peng, L., Jian, S., Kan, Z., Qiao, L., and Li, D. (2024). Not all fake news is semantically similar: contextual semantic representation learning for multimodal fake news detection. *Inf. Process. Manag.* 61:103564. doi: 10.1016/j.ipm.2023.103564

Phan, H. T., Nguyen, N. T., and Hwang, D. (2023). Fake news detection: a survey of graph neural network methods. *Appl. Soft Comput.* 139:110235. doi: 10.1016/j.asoc.2023.110235

Přibáň, P., Hercig, T., and Steinberger, J. (2019). "Machine learning approach to fact-checking in west slavic languages," in *Recent advances in natural language processing [online]*. eds. M. Ruslan, A. Galia (Shoumen, Bulgaria: Incoma Ltd).

Qu, Z., Meng, Y., Muhammad, G., and Tiwari, P. (2023). QMFND: a quantum multimodal fusion-based fake news detection model for social media. *Inform. Fusion* 104:102172. doi: 10.1016/j.inffus.2023.102172

Raja, E., Soni, B., Lalrempuii, C., and Borgohain, S. K. (2024). An adaptive cyclical learning rate based hybrid model for Dravidian fake news detection. *Expert Syst. Appl.* 241:122768. doi: 10.1016/j.eswa.2023.122768

Ruffo, G., Semeraro, A., Giachanou, A., and Rosso, P. (2023). Studying fake news spreading, polarisation dynamics, and manipulation by bots: a tale of networks and language. *Comput. Sci. Rev.* 47:100531. doi: 10.1016/j.cosrev.2022.100531

Soga, K., Yoshida, S., and Muneyasu, M. (2023). Exploiting stance similarity and graph neural networks for fake news detection. *Pattern Recogn. Lett.* 177, 26–32. doi: 10.1016/j.patrec.2023.11.019

Syed, L., Alsaeedi, A., Alhuri, L. A., and Aljohani, H. R. (2023). Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda. *Array* 19:100309. doi: 10.1016/j.array.2023.100309

Tao, W., and Peng, Y. (2023). Differentiation and unity: a cross-platform comparison analysis of online posts' semantics of the russian–ukrainian war based on weibo and twitter. *Commun. Public* 8, 105–124. doi: 10.1177/20570473231165563

Ukrainian news (2024). Kaggle: your machine learning and data science community. Available at: https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news?resource=download (Accessed March 14, 2024).

Xie, B., and Li, Q. (2023). Detecting fake news by RNN-based gatekeeping behavior model on social networks. *Expert Syst. Appl.* 231:120716. doi: 10.1016/j.eswa.2023.120716

Yang, H., Zhang, J., Zhang, L., Cheng, X., and Hu, Z. (2024). MRAN: multimodal relationship-aware attention network for fake news detection. *Comp. Stand. Interf.* 89:103822. doi: 10.1016/j.csi.2023.103822

frontiers | Frontiers in Big Data

TYPE Conceptual Analysis
PUBLISHED 01 November 2024
DOI 10.3389/fdata.2024.1452129

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Dennis Arias-Chávez,
Continental University, Peru
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

*CORRESPONDENCE
Christoph Deppe
christoph.deppe@hsu-hh.de

RECEIVED 20 June 2024
ACCEPTED 16 October 2024
PUBLISHED 01 November 2024

CITATION
Deppe C and Schaal GS (2024) Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept. *Front. Big Data* 7:1452129. doi: 10.3389/fdata.2024.1452129

Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept

Christoph Deppe* and Gary S. Schaal

Helmut-Schmidt-University/University of the Federal Armed Forces, Hamburg, Germany

This study evaluates NATO ACT's cognitive warfare concept from a political science perspective, exploring its utility beyond military applications. Despite its growing presence in scholarly discourse, the concept's interdisciplinary nature has hindered a unified definition. By analyzing NATO's framework, developed with input from diverse disciplines and both military and civilian researchers, this paper seeks to assess its applicability to political science. It aims to bridge military and civilian research divides and refine NATO's cognitive warfare approach, offering significant implications for enhancing political science research and fostering integrated scholarly collaboration.

KEYWORDS

cognitive warfare, disinformation, NATO, hybrid threats, concept analysis

1 Introduction

Cognitive warfare is an emerging academic and military concept that aims to address the exploitation of human cognition and technology to disrupt, undermine, influence, or modify human decision-making [see Claverie and du Cluzel, 2021; du Cluzel, 2021; Reding and Wells, 2022; Deppe, 2023]. It has become increasingly relevant in the current security environment, where adversaries continuously seek to undermine the integrity of political processes in democratic societies, as well as their military-strategic objectives, by deploying sophisticated strategies through coordinated political, military, economic, and information efforts [see Backes and Swab, 2019; Splidsboel Hansen, 2021; Hung and Hung, 2022; Bernal et al., 2020; Adlakha-Hutcheon et al., 2023; Miller, 2023]. Cognitive warfare has roots in early strategies of manipulation and deception in political and military contexts, exemplified by tactics from ancient times, such as those of Sun Tzu. Strategies evolved as information dissemination technologies advanced, broadening its focus from decision-makers to entire populations by leveraging psychological operations, propaganda, and most recently the cyber domain. Today, it encompasses the strategic use of neuroscience, behavioral science, and digital technologies to influence and disrupt human cognition, making it a pivotal element of modern conflict and strategic competition.

Several definitions of cognitive warfare have been published in the literature. Hung and Hung (2022) offer a minimalist conceptualization of cognitive warfare, arguing that it is not a standalone concept but rather a subordinate concept of hybrid warfare and entangled with other traditional non-kinetic warfare forms, such as information warfare and cyber warfare. Cognitive warfare is hereby distinguished by its distinct focus on cognitive effects. Hung and Hung also conceptualize Chinese cognitive warfare as operations that aim to control others' mental states and behaviors by manipulating environmental stimuli. The Chinese approach to hybrid warfare is also based on the Three Warfares concept, which

is a combination of psychological warfare, public opinion warfare, and legal warfare (Lee, 2014).

Bernal et al. (2020) differentiate cognitive warfare from information warfare by delineating their distinct objectives: whereas information warfare centers on controlling the dissemination of information, cognitive warfare strategically aims to shape and manage the reactions of individuals and groups to that information. Backes and Swab (2019) offer a similar view with a general definition: "Cognitive warfare is a strategy that focuses on altering how a target population thinks—and through that how it acts". In relation to Russian information warfare tactics, Tashev et al. (2019) highlight the relevance of the cognitive domain, which they consider to be the relevant aspect of information warfare. Splidsboel Hansen (2021) also illustrates the relevance of a variety of information operations in Russian efforts to influence the cognitive domain. More generally in the Russian approach to warfare, information operations and other measures aimed at reaching cognitive effects are a small number of possible measure among a larger portfolio which includes "all means of national power" (Bartles, 2016). Miller (2023) seeks to distinguish cognitive warfare from related concepts such as cyberwar, cyber conflict and others by referencing four dimension of harm: physical or psychological harm to humans, damage to physical objects, damage to software and data and lastly harm to institutions. Miller sees cognitive warfare as a form of conflict that inflicts psychological harm on individuals and damages institutions. He positions it as a conflict that operates below the thresholds of conventional war and covert operations, thus introducing the nuanced concept of covert cognitive warfare, which subtly undermines its targets without escalating to open hostilities. This view clashes with other conceptualizations of cognitive warfare, that include activities below and above the threshold of war and also kinetic activities as a means to reach cognitive effects (NATO Allied Command Transformation, 2023). A technical evaluation report on a NATO Science and Technology Organization (STO) workshop, which was aimed to contribute to a common understanding of cognitive warfare narrowed the interdisciplinary perspectives from military and academic backgrounds down to include the following components: "the use of technology enabled tactics, techniques, procedures and tools to influence human decision-making at an individual and/or societal level (…) altering human behavior to align with an adversaries' political, social, economic, or military objectives (…) means (i.e., training, technology, policy) to defend and secure the cognitive battlespace (…) resilience and a whole-of-society perspective" (Adlakha-Hutcheon et al., 2023).

NATO Allied Command Transformation (ACT) is tasked with developing a comprehensive military cognitive warfare concept, which will ultimately be part of NATO doctrine. A final version of the cognitive warfare concept is projected to be finalized in late 2024 (NATO Allied Command Transformation, 2023). The cognitive warfare concept by NATO ACT is the most comprehensive attempt to produce a concept under this term to date. Some existing cognitive warfare conceptualizations are subject to conceptual shortcomings, such as unclear definitions, overspecification, conceptual stretching, concept travel etc., which are also due to different conceptual objectives. A unified understanding of a scientific cognitive warfare concept has not yet been reached. This is attributable in part to the nascent nature of the field, but also to the complexity of this interdisciplinary subject area, compounded by numerous technological innovations critical to cognitive warfare. Furthermore, the recognition of a sixth domain of warfare is a different discourse that continues to evolve (Allen and Gilbert, 2010). Because of a number of evolving challenges in global security and information environments there is an ongoing discussion whether a cognitive or human domain could be a sixth domain of warfare (Le Guyader, 2022). However, while notable differences between existing conceptualizations exist, for the purpose of this work, also drawing from proposed NATO definitions, the following working definition of cognitive warfare is used: Cognitive warfare is a tactic, which combines traditional and emerging technologies as well as measures above and below the threshold of war to achieve cognitive effects in an adversary's population, as well as in their political and military leaders. The definition builds upon three key attributes, that need to be present to classify a given attack as cognitive warfare. First, the attack needs to seek cognitive effects, meaning an attempt to alter the cognition of the targets must be present. Second, an element of warfare, meaning a hostile power competition with covert or overt measures above or below the threshold of war. And third, the use of technological means to amplify and/or enable cognitive attacks and their effects. It might be especially relevant for research applications in the sciences, that the attribute technology may also serve to functionally distinguish cognitive warfare from neighboring concepts like hybrid warfare.

It is important to note that the cognitive warfare concept within NATO has been developed as a military concept, to fulfill specific functions within NATO doctrine. Also, the practice and aim of military concept development in NATO differs greatly from concept development in the social sciences (NATO Allied Command Transformation, 2021). While military concepts usually describe new capabilities in the military context, social science concepts aim to produce analytically valuable building blocks, which can be connected to existing theories and be integrated in feasible research designs. The term "cognitive warfare" appears extensively in scientific literature, occasionally in the context of analyzing contemporary conflicts. However, definitions and conceptualizations of cognitive warfare vary significantly across these publications. A critical observation of the existing literature reveals a distinction between conceptualizations of cognitive warfare in scientific literature and contributions in the context of NATO. In the first body of literature, cognitive warfare is frequently viewed as a narrower concept, often subsumed under broader categories such as hybrid warfare. In the latter, NATO's conceptualizations tend to portray cognitive warfare as a comprehensive, standalone concept with broader strategic implications. Therefore, examining NATO ACT's exploratory concept of cognitive warfare will notably improve mutual intelligibility between the different strands of research.

NATO ACT's cognitive warfare exploratory concept represents the most thorough effort to date in formulating a cognitive warfare framework, incorporating contributions from a wide array of both military and civilian researchers within the context of the NATO Science & Technology Organization (STO). Given the extensive collaborative effort behind this concept, it warrants a thorough

examination to assess its relevance to the social sciences and to determine whether, despite potential immediate limitations, it offers any value for future research and practical applications. This study also aims to reduce the conceptual unclarity between military and academic conceptualizations of cognitive warfare and to enhance the development of the concept within both fields of application. Additionally, it seeks to delineate the relationship of cognitive warfare more precisely to adjacent concepts such as FIMI and hybrid threats or hybrid warfare. The lack of clearly defined intensions and extensions of the concept may hinder the understanding of their functional differences, analytic capacity, and interoperability (Sartori, 1970). Therefore, it is of utmost importance to be aware of conceptual mismatches between academic and NATO (military) concepts. The later could hamper the above-mentioned superiority resulting from the connection between academia and NATO. Furthermore, concepts such as cognitive warfare are significant not only in scientific and military contexts but also in political communication. It is probable that cognitive warfare will be employed to engage with the political and public spheres, explaining and justifying research, military actions, and ultimately policy decisions. This pattern has previously been observed with the concept of hybrid threats.

This leads to the following research questions: What are the analytical strengths and weaknesses of the cognitive warfare exploratory concept by NATO ACT from a political science standpoint, and how can cognitive warfare be integrated into existing academic and political conceptual landscapes?

2 Cognitive warfare—Background and organizational considerations

Over the past two decades, the delineations between peace and conflict have become increasingly indistinct. In the same period, a rapid change fueled by technological innovation has radically transformed the way individuals, groups, institutions, and whole societies communicate, how they produce and consume information. The effects of hybrid tactics, such as disinformation, can be observed in many democratic societies. Many activities are attributed to Russia and China (Splidsboel Hansen, 2021; Hung and Hung, 2022; Hellström et al., 2024). A prominent case of state intervention in democratic proceedings was the Russian involvement in the 2016 U.S. presidential election, favoring Republican candidate Donald J. Trump. The specifics of Russian action during this election were outlined in a report compiled by Special Counsel Robert Mueller (2019). Within the EU the European Union External Action Service (EEAS) has registered a significant prevalence of Foreign Information Manipulation and Interference (FIMI) and disinformation, particularly highlighted by Russia's actions during its invasion of Ukraine. The Kremlin strategically used disinformation to justify its actions and manipulate international opinion, leading to unprecedented EU responses, including sanctions against Russian outlets like RT and Sputnik (European Union External Action, 2023b; Deppe and Schaal, 2022). For instance, the EEAS detected almost 400 FIMI incidents in 2022, demonstrating the extensive and coordinated efforts of foreign actors to undermine EU stability (European Union External Action, 2023b). Another example is the EUvsDisinfo initiative by the EEAS, which has documented over 13,300 cases of pro-Kremlin disinformation as of December 2021, revealing a systematic effort to manipulate public opinion within the EU (European Union External Action, 2021). Additionally, China has been implicated in disinformation campaigns that target the EU, particularly concerning COVID-19, where state-controlled media spread false information to counter criticisms of China's handling of the pandemic (European Union External Action, 2023a).

The emergence of artificial intelligence (AI) heralds significant changes in how information is managed and disseminated. On the positive side, AI systems, readily accessible and operable without specialized knowledge, offer unparalleled access to resources, information, and knowledge, potentially enriching decision-making processes by streamlining complex data analysis. Additionally, the capacity for personalization that AI brings can enhance user engagement with information platforms.

Conversely, the advent of AI introduces several challenges. AI models present a high potential for misuse, notably in the mass production of spurious media and disinformation. This capability could profoundly destabilize information ecosystems. Moreover, the risk of exacerbating digital divides looms large, as entities equipped with advanced AI tools could gain disproportionate influence over public discourse and perceptions. Furthermore, while personalization can improve engagement, it also raises significant concerns regarding privacy breaches, targeted influencing of groups and individuals as well as the creation of echo chambers, thereby limiting exposure to diverse viewpoints.

This transformation, along with changes in economic systems, the global security environment and other factors has made it necessary to develop new academic and military concepts to make sense of and analyze novel forms of competition and conflict situations. These concepts include unconventional warfare, hybrid threats, hybrid warfare, and FIMI, to name a few. More recently, considering drastic changes in the global security environment and the rapid emergence of new threats in the cyber and cognitive domains, the concept of cognitive warfare has been introduced to academic and military discourses. Like with any other newly formulated concept, it has to be made sure, that the emerging cognitive warfare concept adds actual analytic utility to diverse research applications, meaning the concept can actually be used to measure a new phenomenon that is not or incompletely captured by existing concepts.

2.1 The purpose of the cognitive warfare concept within NATO

In the 2022 strategic concept (NATO, 2022) NATO highlights activities by Russia and China as significant threats to the Alliance's security and interests. It points out that Russia employs tactics such as coercion, subversion, aggression, and annexation to establish spheres of influence and direct control. The country utilizes a combination of conventional, cyber, and hybrid means against NATO and its partners. Regarding China, the strategic concept notes that its stated ambitions and coercive policies challenge NATO's interests, security, and values. China employs various

political, economic, and military tools to expand its global presence and assert influence (Hung and Hung, 2022; Aukia and Kubica, 2023). However, it maintains opacity about its strategy, intentions, and military build-up. The strategic concept 2022 states that China's aggressive use of hybrid and cyber operations, along with its confrontational rhetoric and dissemination of disinformation, target Allies and pose a threat to Alliance security.

The strategic concept also recognizes the escalating threat of hybrid tactics (NATO, 2022). These tactics encompass a spectrum of measures, including political, economic, energy, and informational methods, employed coercive to attain strategic goals (Aho et al., 2023). Notably, in the strategic concept 2022 NATO acknowledges that, hybrid operations against Allies can escalate to the level of an armed attack, potentially necessitating the invocation of Article 5 of the North Atlantic Treaty (NATO, 2022, 1949). Consequently, NATO aims to improve its capabilities for preparedness, deterrence, and defense against the coercive application of hybrid tactics, by state or non-state actors. The strategic concept also states that the alliance will maintain its support for partners in countering hybrid challenges, striving for optimal collaboration with relevant institutions like the European Union. From this brief analysis of the strategic concept, one of the most important policy documents within NATO, it can be concluded, that hybrid tactics and related activities are a severe vector to NATO, which demands adequate responses. The NATO Warfighting Capstone Concept (NWCC) is subordinate to the strategic concept and is a military concept that outlines NATO's vision for maintaining and developing its military advantage through 2040 (NATO, 2021). It addresses the changing character of war and power competition and emphasizes the need for NATO to be able to operate in all domains, including space and cyberspace. The NWCC also highlights the importance of interoperability and partnerships with other nations and organizations. The NWCC defines "6 Outs", which are a set of functions that the future Alliance MIoP (Military Instrument of Power) must aspire to outperform to maintain NATO's military advantage. These functions are: out-think, out-excel, out-fight, out-pace, out-partner, and out-last (NATO, 2021). The NWCC posits that by achieving superiority in these domains, NATO will enhance its capability to comprehend and counter potential activities by adversaries, cultivate partnerships, and adjust to evolving situations. The resulting warfare development imperatives are five key areas on which NATO must concentrate to achieve these goals. These imperatives are: cognitive superiority, layered resilience, influence and power projection, cross-domain command, and integrated multi-domain defense (NATO, 2021).

Within the five warfare development imperatives listed in the NWCC, the cognitive warfare concept is most relevant in cognitive superiority. Cognitive superiority describes the ability of NATO to better understand the operating environment and potential adversaries relative to its own capabilities and objectives (NATO, 2021). It involves expanding knowledge and understanding across all domains, enabled by technology, to maximize the ability of military leaders to anticipate, think, decide, and act. The goal is to achieve a cognitive advantage over potential adversaries by building better situational awareness and understanding. Hereby, the cognitive warfare concept within NATO doctrine will serve a 2-fold purpose: to improve the comprehension of evolving threats within the cognitive realm and to lay the groundwork for potential future developments in warfare in the cognitive domain (NATO Allied Command Transformation, 2023). The concept shall provide a unified framework for comprehending and effectively addressing cognitive warfare, outlining its dynamics, mechanisms, and implications for both NATO's warfighting capabilities and cognitive superiority. Its overarching objective is to improve NATO's cognitive resilience, as well as to protect and enhance decision-making capacities.

In essence, the cognitive warfare concept shall enhance NATO's understanding of upcoming cognitive threats, protect cognitive resilience by defining potential impacts, and produce a holistic strategy for mitigating the effects of adversarial cognitive warfare through tactics like education, collaboration, protection, and influence in the cognitive domain. This concept thus fulfills a distinct role as a lower-level military concept, positioned subordinate to the NWCC. In due course, the aspiration is for the cognitive warfare concept to be integrated into NATO's official doctrine (NATO Allied Command Transformation, 2023; Groestad, 2022).

3 The cognitive warfare exploratory concept

In this chapter, the cognitive warfare exploratory concept, published by NATO Allied Command Transformation (ACT), is reviewed from a methodological perspective from the social sciences. As of spring 2024 it is the latest published non-draft version of the NATO ACT cognitive warfare concept. In this analysis, focus is placed solely on elements of the exploratory concept that are analytically valuable, omitting detailed discussions of technological, legal, or organizational specifics.

3.1 Basic concept and definitions

The review of the cognitive warfare concept begins at the highest level, with the concept term and the basic definition. The proposed definition of cognitive warfare as published in the exploratory concept by NATO ACT is "Activities conducted in synchronization with other Instruments of Power, to affect attitudes and behavior by influencing, protecting, or disrupting individual and group cognition to gain advantage over an adversary" (NATO Allied Command Transformation, 2023).

The substantive necessity for a concept like cognitive warfare is derived from observed challenges for NATO, which can be broken down into two developments. First, technological progress and shifts in information consumption, wield adversaries' greater capacity to amass and manipulate data, sway emotions, and shape beliefs and behaviors. This enables the leveraging of societal divisions using novel technologies [e.g., artificial intelligence (AI), emerging & disruptive technologies, data harvesting], and the proliferation of social media influencing individuals' thoughts, emotions, and actions. As the authors themselves state, this mirrors hybrid warfare tactics, where adversaries target society as the vector to exert influence indirectly on key targets: political and military leaders. The effectiveness of influence campaigns is conceptualized

to hinge on the calculated manipulation of emotions and cognitive predispositions to instigate widespread shifts in attitudes and behavior. These alterations are frequently nuanced, blurring the distinction between genuine societal debate and discord, and the hostile exploitation of societal divisions through cognitive attacks (NATO Allied Command Transformation, 2023).

Second, it is stressed that cognitive attacks are not new, however the concept defines them as deliberate offensive maneuvers aimed at influencing perceptions, beliefs, interests, decisions, and behavior by directly targeting the human mind. It is conceptualized that the innovation lies in the adversaries' ability to rapidly and anonymously carry out cognitive attacks within the Information Environment (IE) using digital platforms and emerging disruptive technologies (EDTs) (NATO Allied Command Transformation, 2023).

The focus on cognition and the human mind distinguishes the cognitive warfare concept from other concepts like hybrid threats, FIMI and disinformation, whose conceptualized effects often end at the acceptance or rejection of specific information or narratives. In cognitive warfare, the main effect of activities is conceptualized to lie in the manipulation of emotional and subconscious processes of the human mind. This is much more far-reaching than other related concepts like Foreign Information Manipulation and Interference (FIMI) or hybrid threats. The authors clarify that, "synchronized and coordinated attacks on emotions, thoughts and behaviors impact will, morale, decision-making and situational understanding" (NATO Allied Command Transformation, 2023). Furthermore, "Cognitive attacks are designed to use information to activate the subconscious processes in our brains, making it difficult for our conscious minds to perceive the presence of a cognitive threat" (NATO Allied Command Transformation, 2023). These factors constitute an approximation to the empirical referents that partly constitute the extension of the cognitive warfare concept.

3.2 Problem space

The problem space section of the exploratory concept provides an overview of the potential dangers posed by cognitive warfare, again adding to the empirical referents of the concept. It begins by labeling cognitive warfare to be a value-neutral set of tactics, which may be employed at every stage on the continuum of competition. It is furthermore problematized as a Whole-of-Society Problem in which "adversaries are targeting the NATO Alliance through campaigns to malignly influence the attitudes, decisions and behaviors of individuals, groups and societies. Emerging and Disruptive Technologies (EDTs) and sciences enable these cognitive attacks. Our adversaries aim to turn our strengths into vulnerabilities that weaken the Alliance" (NATO Allied Command Transformation, 2023). In this definition the role of technological innovations in the distribution of influence campaigns is highlighted as a powerful enabling factor, that can be used to attack the discourse spaces of open liberal democratic societies. Furthermore, the military challenge of cognitive warfare is described as "Alliance decision-making, mission and forces are directly and indirectly vulnerable to cognitive attacks. The role of the Military Instrument is the cognitive dimension is unclear, particularly below the threshold of armed conflict. This causes gaps in policy, defense planning and capabilities" (NATO Allied Command Transformation, 2023). Hereby, the concept is connected to ongoing discussions in many democratic societies, about the role of different governmental institutions in the mitigation of threats in the information environment.

Next, the kind of actions, that are considered to be part of cognitive warfare are listed. These tactics, or vectors and enablers can be the defining attributes that constitute the meaning or intension of the cognitive warfare concept[1]. Cognitive attacks, both presently and potentially in the future, are described to be facilitated through a range of vectors, capabilities, and enablers:

Traditional vectors and enablers:

This category includes kinetic force and established channels like broadcast and print mass media. Additionally, it involves various actors such as corporate, state, and political entities, along with interpersonal engagement (NATO Allied Command Transformation, 2023).

Existing technology vectors and enablers:

This domain leverages contemporary technology. It encompasses social media platforms, the utilization of big data, the integration of augmented reality and wearable smart devices, as well as the use of gaming and encrypted communication platforms. Avatars and virtual profiles are also instrumental in this context (NATO Allied Command Transformation, 2023).

Emerging technology vectors and enablers:

This category delves into cutting-edge technologies that hold significant potential for cognitive attacks. It encompasses synthetic media, exemplified by deepfakes and AI-driven media. Additionally, it includes the widespread use of artificial intelligence, the immersive realm of the Metaverse, and the concerning emergence of neuroweapons. The listed vectors and enablers encompass a wide spectrum of tactics and approaches, underscoring the wide intension of the cognitive warfare concept (NATO Allied Command Transformation, 2023).

Further, the authors identify various individual risk factors and resulting triggers that heighten susceptibility to micro-level cognitive attacks. These include deficiencies in accurate knowledge, deeply ingrained worldviews, negative emotional experiences, and limited literacy (NATO Allied Command Transformation, 2023). Addressing these factors is considered crucial for enhancing the resilience of both NATO personnel and member nations against cognitive attacks. This can be achieved by improving knowledge, critical thinking, and emotional resilience. Additionally, the document outlines three primary triggers influencing vulnerability to influence and manipulation. These are cognitive inflexibility, the

1 Annex C of the NATO ACT cognitive warfare exploratory concept is an attempt at an actor and value neutral taxonomy of cognitive warfare. However, as the authors themselves state, a significant overlap between different measures exists. Since the taxonomy offers little analytical potential, it is not subject to analysis in the present work. Furthermore, a highly differentiated taxonomy of an already extensive concept like cognitive warfare poses a high risk of an over specification of the concept.

need for social belonging, and emotional arousal. Mitigating these triggers is vital for bolstering resilience against cognitive attacks. This involves promoting cognitive flexibility, fostering a sense of belonging, and managing emotional arousal.

The authors also name risk factors that heighten vulnerability to cognitive attacks on the meso-level i.e., in social and cultural groups. These factors include group polarization and social trust (NATO Allied Command Transformation, 2023). Group polarization, influenced by human social tendencies, can be exacerbated by social media platforms through algorithms that reinforce existing beliefs. This fosters confirmation bias and the spread of disinformation, undermining trust and manipulating groups. Social trust, crucial for societal cohesion, can be exploited by malign actors spreading disinformation to erode trust in institutions and leaders. Addressing these risk factors is conceptualized to be crucial for enhancing the resilience of NATO personnel and member nations against cognitive attacks. This may be achieved by promoting critical thinking, bolstering social trust, and countering group polarization.

Finally, the authors also examine risk factors on the macro level, i.e., societies and nations. The concept highlights nations' varying susceptibility to cognitive attacks, a crucial consideration for NATO in evaluating member nations' resilience (NATO Allied Command Transformation, 2023). While NATO's MIoP does not possess a direct mandate to address these factors, their understanding remains essential. They are conceptualized as the basis for collaborative efforts with NATO partner nations and non-NATO organizations.

In liberal democracies, a notable vulnerability exists to adversarial cognitive warfare, challenging NATO's foundational values (NATO Allied Command Transformation, 2023) as well as democratic core values. While preserving the liberal democratic system remains a priority for NATO, it is recognized as a risk factor due to the principled rejection of authoritarian control methods. Adversaries perceive this as a significant vulnerability that can be exploited to sow discord within societies and erode the ability to govern in line with liberal democratic principles (Deppe, 2023).

Furthermore, the authors name information and media literacy as an important risk factor, as research indicates a direct correlation with susceptibility to disinformation and cognitive manipulation (NATO Allied Command Transformation, 2023). Citizens, including NATO personnel, often remain unaware of their own vulnerabilities to cognitive manipulation. Therefore, the authors underscore the necessity for heightened information and media literacy efforts to counter cognitive manipulation attempts.

Civic engagement encompasses activities that enhance community wellbeing through political and non-political means, breaking down barriers and augmenting societal resilience (NATO Allied Command Transformation, 2023). While improving civic engagement falls beyond NATO's political and military scope, recognizing its protective potential offers an opportunity for the Alliance to collaborate with external entities focused on fortifying societal resilience.

The authors state that, NATO has observed a surge in anti-establishment populism, indicating discontent with prevailing economic, social, and cultural conditions in numerous NATO member nations (NATO Allied Command Transformation, 2023). In some instances, the growing support for populist ideologies may also signify the influence and success of cognitive attacks by adversaries. These are achieved through various means, including espionage, hacking, disinformation campaigns, and covert funding of political movements.

From a methodological viewpoint, the risk factors for an increased vulnerability to cognitive attacks on the micro, meso and macro levels, listed above, can potentially be attributed to be part of the extension of the concept. This is because these factors can be read as a list of variables, that constitute a case, that would be captured by the concept cognitive warfare. From the counter perspective, if a given hypothetical case featured none of the listed risk factors, it would not be captured by a measurement that captures instances of cognitive warfare.

In a subsequent section, the exploratory concept provides a list and description of the intended effects of cognitive warfare (NATO Allied Command Transformation, 2023). Cognitive warfare is conceptualized to entail a diverse range of intended effects, posing intricate challenges in recognizing attacks and their protracted consequences. In this context, cognitive attacks are employed within broader geopolitical strategies to hinder decision-making processes, erode national or institutional unity, sow societal division, exploit identities and narratives, and undermine the resolve to engage in conflict.

First, under "Impede Decision-Making and Disrupt OODA Loop," (NATO Allied Command Transformation, 2023) decision-making, contingent on information availability, becomes susceptible to manipulation by state and non-state actors. Disinformation compounds uncertainty or propagates false narratives, thereby influencing decision-makers across strata. As an illustration of efforts to undermine decision-making through disinformation, the authors cite Russia's Reflexive Control (RC) theory. The RC theory is conceptualized with the objective of impeding NATO's decision-making processes.

Second, "Divide and Polarize Society" (NATO Allied Command Transformation, 2023) pertains to deep-seated societal polarization, imperiling democracy. Adversaries exploit disinformation to systematically erode social trust, weaken institutions, and impede efforts to reconcile conflicting values and interests. This leads to societal segmentation based on various criteria.

Third, "Weaponize Identity" (NATO Allied Command Transformation, 2023) emphasizes the pivotal role of identity in cognitive warfare, influencing connections to others, societal roles, and cultural and national affiliations. Understanding the potential weaponization of identity is crucial, especially when safeguarding NATO personnel against targeted cognitive attacks.

Next, "Weaponize Narratives" (NATO Allied Command Transformation, 2023) highlights how historical memory and heritage significantly shape individual and group identities, influencing the narratives employed to depict how individuals, communities, and nations perceive themselves. Adversaries adeptly manipulate, discredit, and alter narratives to align with their strategic objectives.

Last, "Impact the Will to Fight" (NATO Allied Command Transformation, 2023) underscores that effective cognitive warfare

requires seamless synchronization and coordination to manipulate human cognition, influencing decision-makers' comprehension of the information environment and their resolve to engage in conflict. Cognitive attacks introduce friction within military leadership, potentially eroding trust in NATO leadership and the overarching alliance mission among military personnel over time.

The intended effects of cognitive warfare listed above can best be attributed to the intension of the concept, because they illustrate what an adversary seeks to accomplish by employing measures and tactics conceptualized as part of cognitive warfare. Analytically, "intension" is very difficult to operationalize, however, it is frequently used in related concepts like disinformation (see Wardle and Derakhshan, 2017) or FIMI (see European Union External Action, 2023a). This concludes the review of areas within the exploratory concept that are relevant to the concept specification of cognitive warfare from a methodological perspective.

4 Concept analysis

4.1 Methodology and data

The definition, use, and significance of concepts in social science is a complex and contested research area (Sartori, 1970, 1984; Collier and Mahon, 1993; Gerring, 2001). In comparative research within the social sciences and beyond, clear and comprehensible concepts are paramount to ensure communicability and intelligibility (Sartori, 1984). In its core form, a referential concept consists of a term, that names the concept; one or more empirical referents, that are captured by the concepts, thereby defining the denotation or extension of a given concept; and lastly, one or more defining attributes, that fill the concept with meaning, defining the connotation or intension of a given concept (Sartori, 1984; Gerring, 2001). Concepts serve as the foundation of the scientific process, informing research questions and hypotheses; they are essential to the development of research designs and to many downstream tasks of research, such as operationalization and research communication. The definition of concepts represents a fundamental preliminary step in the planning of a research endeavor. "When a concept is formulated (or reformulated) it means that one or all of the features is adjusted. Note that they are so interwoven that it would be difficult to change one feature without changing another. The process of concept formation is therefore one of mutual adjustment" (Gerring, 2012). In order to improve the integration and analytic capacity of concepts in applied research, Gary Goertz has introduced to basic concept model. This is a complex structure that enables the use of complex concepts, considering their multidimensional and multilevel properties (Goertz, 2002). The basic concept consists of three levels. First, the basic level, which is the concept identifier as used in hypotheses and theories. Second is the secondary level which includes the concepts defining features or attributes. Third is the data indicator level, which describes what specific data is indicative of the presence of a given attribute. Items on level three are therefore indicative of items on level two, whereas level one and level two share an ontological relationship (Goertz, 2002). The items on the levels two and three can be aggregated in different modes to suit concept meaning and measurement considerations. Goertz (2002) notes that the model can be read top-down when referring to the conceptualization and semantics of a given model, as well as bottom-up when considering the measurement and numerics. The strengths of the model lie in making complex theoretical concepts measurable. The basic concept model is therefore used below produce a concept of cognitive warfare, based on the concept by NATO ACT which can be applied in the social sciences[2]. However, for the analysis of the cognitive warfare concept, the referential concept model will be used. This is because, the cognitive warfare concept by NATO ACT is a very extensive maximalist concept with a military background that cannot be directly reformatted into the format of the basic concept model without some amount of reconceptualization. However, this would distort the concept analysis. Therefore, the uncomplicated referential concept model is suited much better for a concept analysis in this case since the assignment of elements of the analyzed concept can be allocated to components of the referential concept model is much more straightforward.

Because the quality of concepts is critical to many research endeavors, criteria for the evaluation of concepts have been developed. Sartori's (1970) approach emphasizes the importance of conceptual clarity and precision, advocating for the use of a "ladder of abstraction" to avoid concept stretching in comparative politics. Building on this, Collier and Mahon (1993) refined Sartori's model by offering systematic methods for adjusting conceptual categories, ensuring validity across diverse contexts.

While Goertz's (2002) multi-dimensional model provides a detailed structure for defining complex concepts, it is not suitable for the present conceptual analysis due to its rigid conceptual hierarchy. Adcock and Collier (2001) propose a model that emphasizes the integration of qualitative and quantitative approaches to achieve both content and measurement validity, ensuring that concepts are appropriately operationalized for empirical research. They focus on aligning theoretical definitions with measurable indicators to maintain conceptual rigor.

Transitioning from this, Gerring's model offers a specific set of criteria, such as differentiation, coherence, and utility, that guide the evaluation of concepts, helping to tailor the conceptual framework to the unique needs of my research. Gerring's model is particularly well-suited for the analysis of complex, maximalist concepts due to its comprehensive framework for assessing coherence, differentiation, and utility across diverse contexts. This approach emphasizes the internal structure and definitional clarity of a concept, which is essential for navigating the intricacies and multiple dimensions inherent in maximalist constructs. By focusing on these criteria, Gerring's model facilitates a systematic evaluation that is well-aligned with the complexities typically associated with such expansive conceptual frameworks. In contrast, the model by Adcock and Collier is more focused on ensuring measurement validity and bridging qualitative and quantitative methods, but it may not provide the same depth of analysis needed to unpack and assess the intricate dimensions of a highly complex concept. In his highly cited works on social science methodology Gerring (2001, 2012) has published two frameworks to evaluate concepts: "Criteria of conceptual goodness" as well as "Criteria of conceptualization".

2 See Section 5.

For the application of the evaluation of an already existing concept, the "Criteria of conceptual goodness" (Gerring, 2001) are most suitable.

Gerring (2001) outlined eight criteria for evaluating conceptual goodness: coherence, operationalization, validity, field utility, resonance, contextual range, parsimony, and analytical/empirical utility. Coherence ensures that the concept is internally consistent and free from contradictions, while operationalization emphasizes the ease with which the concept can be translated into measurable indicators for empirical analysis. Validity refers to the extent to which the concept accurately captures the phenomena it is intended to describe. Field utility concerns the concept's practical relevance and usefulness to the field, whereas resonance gauges how intuitively and broadly the concept aligns with existing knowledge or understanding in the discipline. Contextual range assesses the concept's applicability across different settings, ensuring it is versatile without losing its meaning. Parsimony encourages simplicity, aiming for the concept to convey essential ideas without unnecessary complexity. Finally, analytical/empirical utility evaluates the concept's ability to generate meaningful, testable hypotheses and contribute to both theoretical and empirical advancements.

These eight criteria will serve as a comprehensive guideline for evaluating the cognitive warfare exploratory concept as published by NATO ACT (NATO Allied Command Transformation, 2023). By applying each criterion to the concept, a thorough and structured assessment can be conducted, ensuring that the concept's internal structure, practical relevance, and empirical applicability are fully examined. This process aims to provide a coherent evaluation of the cognitive warfare concept's overall quality and utility in the context of scientific research to aid in making the miliary concept of cognitive warfare useable in scientific applications.

4.2 Concept evaluation

To methodologically evaluate the cognitive warfare concept, the learnings of the concept review in Section 3 will be assigned to the basic elements of a referential concept. At its core, a concept includes three elements: the term or linguistic label itself, the extension or empirical referent, and the intension, i.e., the defining attributes of a given concept which fill the concept with meaning (Sartori, 1984; Gerring, 2001; Wonka, 2007).

The term of the concept is cognitive warfare, thereby clarifying that the concept describes a type of warfare happening in the cognitive domain/dimension. The defining attributes, that fill the concept with meaning (intension or connotation) are divided in two categories. The first set of attributes that constitute cognitive warfare are its operations and tactics, which are traditional vectors and enablers, existing technology vectors and enablers as well as emerging technology vectors and enablers. The second set of attributes are the intended effects of cognitive warfare, which are impeding decision-making and the disruption of the Observe, Orient, Decide, Act (OODA) loop, the division and polarization of societies, weaponizing identity, weaponizing narratives and impacting the will to fight.

The concept's extension, or its empirical referents, poses a greater challenge to apprehend compared to its defining attributes. This complexity arises from the evolving nature of cognitive warfare and the nascent stage of many technologies outlined in the exploratory concept. As a result, pinpointing precise instances of cognitive warfare proves problematic; cases of cognitive warfare in the contemporary scientific literature often follow a radically different conceptualization of cognitive warfare (see for example 7). Consequently, a theoretical scenario of cognitive warfare is inferred from the exploratory concept. Here it can be deduced that the extension of the concept could be a series of synchronized cognitive attacks, defined as "offensive actions employed to achieve effects on perceptions, beliefs, interests, aims, decisions and behaviors by deliberately targeting the human mind" (NATO Allied Command Transformation, 2023) in the information environment, using EDTs, in individuals, groups or societies, which are particularly vulnerable to cognitive attacks.

After assigning the elements of the cognitive warfare exploratory concept to the basic elements of a referential concept, the next step is its evaluation using Gerring's (2001) eight criteria for conceptual goodness. The first criterion is coherence, which inquires how internally coherent and externally differentiated a concept's attributes are regarding neighboring concepts. For cognitive warfare, the internal coherence can be considered high because the different attributes build upon each other to characterize the clearly defined mechanisms. The cognitive warfare exploratory concept meticulously outlines several measures and effects that are considered to be cognitive warfare, in sum forming a coherent concept. As far as the external differentiation goes, the concept differs from neighboring concepts in several key issues, namely the focus on cognitive effects and actions in the information environment using EDTs, and the sector specific focus on the military and the protection of the MIoP, the concept is therefore sufficiently coherent. However, it could be argued that many attributes associated with cognitive warfare might equally apply to adjacent concepts such as hybrid threats or hybrid warfare, posing an analytical challenge. This issue could be resolved in two ways. Firstly, by acknowledging that cognitive warfare may be too akin to the established concepts of hybrid threats and hybrid warfare, implying it offers limited analytical utility or merely serves as a subordinate tactic within these broader concepts. Alternatively, the cognitive warfare concept could undergo reconceptualization, emphasizing those features that distinctly set it apart from related concepts. This would likely entail a sharper focus on Emerging Disruptive Technologies (EDTs) and cognitive processes and effects that extend beyond well-known strategies such as disinformation, propaganda, and information operations.

The second criterion, operationalization, probes the concept's ability to differentiate its own referents from other empirical referents distinctly. In this regard, the cognitive warfare concept faces a number of challenges. First, as the concept is concerned with emerging technologies, which is on reason why it might currently lack concrete instances in the field. Second, the concept is inherently future oriented, which means that is not aimed to measure present day instances. Lastly, due to the covert nature of many tactics associated with cognitive warfare, detecting its occurrence may be challenging, even if they were to happen. Classifying a conflict or power competition as cognitive warfare

involves assessing whether it meets specific thresholds. This approach is consistent with the measurement of many complex concepts in the social sciences, where determining the presence of a phenomenon often requires evaluating multiple dimensions. For measuring cognitive warfare, it could be argued that some present conflicts, like recent Russian activities in the Baltic (Oksanen et al., 2024), Chinas behavior toward Taiwan and its activities in the Taiwan Strait (Hung and Hung, 2022), Chinas activities in the South China Sea dispute (Hong, 2013) as well as the Russian warfare in its war on Ukraine and the Ukrainian defensive effort (Muradov, 2022; Dov Bachmann et al., 2023) could be classified as cognitive warfare, provided all essential attributes of the concept are present. From a quantitative perspective, the challenge lies in determining whether the intensity and scale of these observed activities sufficiently meet the thresholds required to identify each necessary attribute of cognitive warfare in these specific cases.

The third criterion, validity, addresses whether the concept accurately measures what it is intended to represent. This evaluation is challenging given the difficulties of classifying current conflicts or power competitions as cognitive warfare with a maximalist concept like the NATO ACT concept. It is pertinent to note once more, that the concept is future-oriented, and present-day instances would either be classified under different conceptual frameworks or the threshold of classifying a given case as cognitive warfare would need to be adjusted. Therefore, it can be concluded that identifying and measuring concrete instances of cognitive warfare using the NATO ACT concept in the field is challenging. This may change with further advancements in EDTs, which play in important role in cognitive warfare. However, a slight reconceptualization can change this outlook [2].

The fourth criterion, field utility, assesses the practical usefulness of the concept in comparison to similar ones. Currently, in the realm of cognitive warfare, concepts like hybrid threats, hybrid warfare or information warfare hold greater analytical utility. However, as technological capacities continue to advance, cognitive warfare has the potential to offer a significant contribution in comprehending forthcoming threats more effectively. The concept does indeed describe a novel, emerging threat, which has previously not been described by existing concepts.

The fifth criterion, resonance, examines whether the concept holds relevance in both general and specialized contexts. Within NATO and military circles, the concept of cognitive warfare finds resonance, because of its function in official NATO doctrine. However, in broader non-military contexts, the explicit military focus and the use of the term "warfare" in cognitive warfare may complicate communication efforts. Established concepts like hybrid threats are more commonly used in these scenarios, particularly because terms incorporating "warfare" tend to be less palatable to the general public. However, the concept might be an effective tool to analytically focus analyses on cognitive effects in a diverse set of adversarial measures. While the cognitive warfare concept by NATO ACT is designed to be a concept that shall help to develop tactics to defend NATO against potential cognitive warfare by adversaries, a potential drawback of the term cognitive warfare could also be, that the term could be misperceived in the general public or even misused in adversarial disinformation.

The sixth criterion, contextual range, assesses the concept's applicability across different languages. "cognitive warfare" is a term that distinctly conveys its meaning and can be meaningfully translated. However, in other strategic cultures other concepts exist, such as Russian reflexive control (Jaitner and Kantola, 2016) and Chinese indirect approaches (Aukia and Kubica, 2023) or the three warfare strategy (Lee, 2014), which partially overlap with cognitive warfare, making the concept less meaningful in these cultural contexts.

The seventh criterion, parsimony, evaluates the conciseness of the term and its list of attributes. The term "cognitive warfare" itself is succinct and precise. However, its attributes in the NATO ACT concept are extensive and complex, as they are feature an long list of measures and effects, which form a complex maximalist concept. The concept includes many measures and effects because it is a military concept, which must be linked to specific defense capabilities. For a scientific application the concept could benefit from a reduced list of attributes, which focuses on the most important measures and effects, while removing some attributes which are analytically less relevant.

The eighth criterion, analytic/empirical utility, pertains to how useful the concept is in analytic contexts and research designs. In contexts focused on emerging threats, "cognitive warfare" holds significant analytical potential. It could serve as a tool to grasp and conceptualize several possible cognitive effects of many adversarial measures in conflicts below and above the threshold of war. Nevertheless, the concept's applicability in contemporary empirical research remains constrained. This limitation arises from the broad array of attributes, the emphasis on emerging disruptive technologies (EDTs), and the necessity for the concurrent application of various measures. Collectively, these factors establish an exceedingly high benchmark for classifying a scenario as cognitive warfare, rendering the concept impractical for empirical analysis.

5 Discussion

The evaluation of the cognitive warfare concept based on John Gerring's eight criteria for conceptual goodness reveals several insights. First the lack of real-world instances of cognitive warfare as defined by NATO ACT complicates the validation of the concept and its current practical value, yet its forward-looking nature is in keeping with its intended purpose. Although hybrid threats and hybrid warfare currently offer greater practical utility, cognitive warfare is expected to become increasingly pertinent as technological advancements continue, thereby enhancing its future relevance in the field.

As mentioned above, the term "cognitive warfare" is clear and translatable across languages. However, the term may be difficult in different strategic cultures. Also, the introduction of a concept with the term "warfare" in the title may be problematic in some political and societal arenas. While concise, the list of attributes may require further elaboration. Lastly, in contexts focused on emerging threats,

the concept holds significant analytical potential, though its current empirical applicability may be limited.

The concept demonstrates high internal coherence, as its attributes synergistically characterize defined mechanisms. However, the external differentiation from neighboring concepts can potentially be problematic. The concepts unique features lie in analyzing potential cognitive effects through the use of EDTs, and its specific focus on the military sector and the protection of the MIoP. In the existing conceptual landscape, it is debatable whether cognitive warfare needs to be defined as a standalone concept. Instead, it could be more effectively conceptualized as a subordinate concept—or tactic—within the frameworks of hybrid threats or hybrid warfare. This approach could address many of the conceptual challenges associated with cognitive warfare, particularly those related to extensive lists of attributes and conceptual stretching.

Furthermore, the concept faces challenges in operationalization, as concrete instances in the field are currently lacking. This is partly due to the extensive list of defining attributes and an inherent focus on emerging technologies. Furthermore, the covert nature of many associated tactics makes detection difficult. In the literature, more streamlined conceptualizations of cognitive warfare have been introduced. This presents a trade-off: while leaner concepts are generally more suitable for social scientific research and political communication, they may offer limited utility in military contexts. This is because such conceptualizations describe fewer capabilities and might omit critical attack vectors. A practical approach would be to develop a smaller, more focused concept that is interoperable with the broader, more comprehensive frameworks. This conceptualization of cognitive warfare would therefore need to be interoperable with the expansive military concept of cognitive warfare by NATO ACT, while also aligning with contemporary literature on hybrid threats or hybrid warfare.

Above we have introduced a working definition of cognitive warfare, which builds upon existing literature and the concept by NATO ACT: *Cognitive warfare is a tactic, which combines traditional and emerging technologies as well as measures above and below the threshold of war to achieve cognitive effects in an adversary's population, as well as in their political and military leaders*. In order to demonstrate how the cognitive warfare concept by NATO ACT could be reconceptualized be gain more analytical value in a scientific application by streamlining it, we visualize our working definition of cognitive warfare, using the basic concept model by Goertz (2002).

Level 1, or the basic level in Figure 1, represents the core concept of cognitive warfare. As previously discussed, our working definition of cognitive warfare includes three essential elements: achieving cognitive effects, elements of warfare, and the utilization of technology. This basic level outlines the broad, foundational idea of cognitive warfare, serving as an entry point into the concept's more detailed structure.

At Level 2, or the secondary level, we see the breakdown of the concept into its core attributes. These attributes are ontologically interconnected, meaning they work together to give the concept its specific meaning and function. For our working definition, the three attributes—cognitive effects, warfare elements, and technology—are necessary conditions; all must be present in any given case to classify it as an instance of cognitive warfare. The attributes at this level offer a more precise framework for understanding the internal structure of the concept.

Level 3, the data indicator level, focuses on the empirical indicators that suggest the presence of the attributes outlined in Level 2. Here, the concept adopts what Goertz (2002) refers to as a "mean concept structure," meaning that not every indicator needs to be present for an attribute to be identified. Instead, a subset of indicators may suffice, as long as the cumulative evidence surpasses a predetermined threshold. In other words, while the concept's attributes are necessary, the empirical evidence for each attribute need not be exhaustive—only sufficiently robust to suggest the presence of the attribute. This tiered structure allows for a nuanced application of the cognitive warfare concept, where the intensity and combination of indicators at Level 3 guide the assessment of whether a given case qualifies under the overarching framework at Level 1.

Our working definition of cognitive warfare exemplifies how a concept can be reconceptualized to enhance its value for scientific applications while maintaining its core integrity and interoperability with NATO ACT's framework. By focusing on three essential elements—achieving cognitive effects, elements of warfare, and the use of technology—we streamline the NATO ACT definition, making it more operationalizable and empirically testable without sacrificing the essence of the concept. This refined structure preserves the emphasis on the military sector and cognitive effects while addressing the challenges of its broader, more abstract attributes. Importantly, our reconceptualization maintains compatibility with NATO ACT's approach, ensuring that the concept remains interoperable in both military and strategic contexts, while also being better suited for academic analysis. This balance between clarity, coherence, and operational utility strengthens the concept's relevance for both scientific inquiry and real-world applications.

6 Conclusion

In conclusion, the conceptualization of cognitive warfare presents both challenges and opportunities for military and analytical applications. The analysis reveals that the cognitive warfare concept by NATO ACT is an inherently future oriented concept, which can be classified as a maximalist approach to the conceptualization of cognitive warfare. There is considerable conceptual overlap between cognitive warfare and neighboring concepts such as Foreign Information Manipulation and Interference (FIMI), hybrid threats, and hybrid warfare, given that cognitive warfare shares numerous attributes with these established frameworks. This overlap suggests that cognitive warfare could be effectively integrated as a tactical element within broader hybrid threat strategies, thereby enhancing the analytical depth and understanding of hybrid effects by situating cognitive operations within an established, multifaceted context. The present maximalist conceptualization must undergo a reconceptualization and streamlining to improve its utility in research applications. Any concept resulting from a reconceptualization must be interoperable, aligning with both the comprehensive military

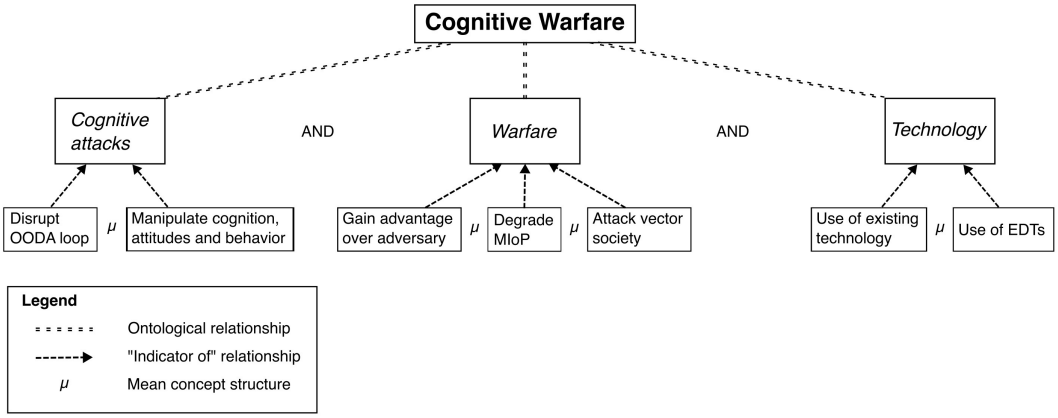

FIGURE 1
Visualization of the cognitive warfare working definition.

perspective of cognitive warfare and the academic insights provided by hybrid threats literature. Above we presented an option how a maximalist cognitive warfare concept could be reconceptualized to make it empirically measurable and improve its analytic utility.

Cognitive warfare has a notable focus on the use of Emerging Disruptive Technologies (EDTs) and the resulting cognitive effects. This innovation highlights the concept's relevance in modern warfare where technological advancements play a pivotal role. However, the empirical analysis of cognitive warfare is currently hampered by a lack of instances in the field and the challenges of measurement. Despite these difficulties, the concept holds considerable potential for analyses focused on foresight and capacity building, where its forward-looking nature can provide significant insights. Furthermore, its distinct focus on cognitive effects could potentially enhance the explanatory power of hybrid threats frameworks when used as a subordinate concept of hybrid threats or hybrid warfare.

Within the NATO doctrine, the cognitive warfare concept will fulfill its function as a lower-level military concept. Given its focus on rapidly evolving EDTs, it is likely to require frequent updates and modifications to remain effective and relevant. This need for continual adaptation speaks to the dynamic nature of cognitive warfare and the fast-paced technological environment in which it operates.

The NATO ACT's conceptualization of cognitive warfare represents a maximalist definition, incorporating a broader range of attributes compared to more minimalist interpretations found in the literature. This broader approach, as developed by NATO ACT, allows for a more comprehensive application in strategic military planning and operations. However, it also necessitates a clear understanding and delineation of cognitive warfare to prevent conceptual stretching and ensure its effective integration into military and analytical frameworks.

By considering these aspects, it becomes evident that while cognitive warfare is a potent and evolving concept, its successful implementation and utility depend on careful consideration of its scope and the context in which it is applied. As we move forward, it will be crucial to continue refining the concept to ensure it remains a valuable tool in the arsenal of modern military strategy and analysis.

Author contributions

CD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. GS: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This publication has been funded by the Open-Access-Publication-Fund of the Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg.

Acknowledgments

Part of this work was presented at the NATO Science and Technology Organization HFM-361 Symposium on "Mitigating and Responding to Cognitive Warfare" 2023 in Madrid, Spain. The authors thank the attendees of the symposium for their valuable comments.

Conflict of interest

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

Publisher's note

References

Adcock, R., and Collier, D. (2001). Measurement validity: a shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.* 95, 529–546. doi: 10.1017/S0003055401003100

Adlakha-Hutcheon, G., Dābakk, S. L., Danley, L., Bērziņš, J., and Blatny, J. M. (2023). *Advancing Towards a Common Understanding of Cognitive Warfare for Science and Technology, and Identifying Future Research Trajectories*. NATO Science & Technology Board.

Aho, A., Alonso Villota, M., Giannopoulos, G., Jungwith, R., Lebrun, M., Savolainen, J., et al. (2023). *Hybrid Threats: A Comprehensive Resilience Ecosystem*. Hybrid CoE - The European Centre of Excellence for Countering Hybrid Threats. Available at: https://www.hybridcoe.fi/publications/hybrid-threats-a-comprehensive-resilience-ecosystem/ (accessed October 26, 2023).

Allen, P. D., and Gilbert, D. P. (2010). Qualifying the information sphere as a domain. *J. Inf. Warfare* 9, 39–50. Available at: https://www.jinfowar.com/journal/volume-9-issue-3/qualifying-information-sphere-domain

Aukia, J., and Kubica, L. (2023). *Russia and China as Hybrid Threat Actors: The Shared Self-other Dynamics*. Helsinki: The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE Research Report). Report No.: 8. Available at: https://www.hybridcoe.fi/publications/hybrid-coe-research-report-8-russia-and-china-as-hybrid-threat-actors-the-shared-self-other-dynamics/ (accessed April 2, 2024).

Backes, O., and Swab, A. (2019). *Cognitive Warfare: The Russian Threat to Election Integrity in the Baltic States*. Harvard Kenney School Belfer Center for Science and International Affairs. Available at: https://www.belfercenter.org/publication/cognitive-warfare-russian-threat-election-integrity-baltic-states (accessed February 3, 2023).

Bartles, C. K. (2016). Getting gerasimov right. *Milit. Rev.* 96, 30–8. Available at: https://research.ebsco.com/linkprocessor/plink?id=1d320849-2dfc-3cf0-95f3-0b2269b96b30

Bernal, A., Carter, C., Singh, I., Cao, K., and Madreperla, O. (2020). *Cognitive Warfare - An Attac on Thought and Truth*. Baltimore MD: Johns Hopkins University. Available at: https://www.innovationhub-act.org/sites/default/files/2021-03/Cognitive%20Warfare.pdf

Claverie, B., and du Cluzel, F. (2021). *The Cognitive Warfare Concept*. NATO ACT Innovation Hub, 11. Available at: https://www.innovationhub-act.org/sites/default/files/2022-02/CW%20article%20Claverie%20du%20Cluzel%20final_0.pdf (accessed January 18, 2023).

Collier, D., and Mahon, J. E. (1993). Conceptual 'stretching' revisited: adapting categories in comparative analysis. *Am. Polit. Sci. Rev.* 87, 845–855. doi: 10.2307/2938818

Deppe, C. (2023). Disinformation in cognitive warfare, foreign information manipulation and interference, and hybrid threats. *Defence Horizon J.* doi: 10.5281/zenodo.10005172

Deppe, C., and Schaal, G. S. (2022). *Vernetzte Desinformationskampagnen: Der Fall Nawalny*. #GIDSresearch. Available at: https://openhsu.ub.hsu-hh.de/handle/10.24405/14603 (accessed November 23, 2022).

Dov Bachmann, S. D., Putter, D., and Duczynski, G. (2023). Hybrid warfare and disinformation: a Ukraine war perspective. *Glob. Policy* 14, 858–869. doi: 10.1111/1758-5899.13257

du Cluzel, F. (2021). *Cognitive Warfare*. NATO ACT Innovation Hub, 45. Available at: https://www.innovationhub-act.org/sites/default/files/2021-01/20210122_CW%20Final.pdf (accessed February 3, 2023).

European Union External Action (2021). *2021 StratCom activity report - Strategic Communication Task Forces and Information Analysis Division*. Bruxelles. Available at: https://www.eeas.europa.eu/eeas/2021-stratcom-activity-report-strategic-communication-task-forces-and-information-analysis_en (accessed August 18, 2023).

European Union External Action (2023a). *1st EEAS Report on Foreign Information Manipulation and Interference Threats*. Bruxelles: European Union External Action. Available at: https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en (accessed February 5, 2024).

European Union External Action (2023b). *2022 Report on EEAS Activities to Counter FIMI*. Available at: https://www.eeas.europa.eu/eeas/2022-report-eeas-activities-counter-fimi_en (accessed February 5, 2024).

Gerring, J. (ed.). (2001). "Concepts," in *Social Science Methodology: A Criterial Framework* (Cambridge: Cambridge University Press), 33–64.

Gerring, J. (2012). *Social Science Methodology: A Unified Framework, 2nd Edn*. Cambridge: Cambridge Univ. Press, xxv+495.

Goertz, G. (2002). *Social science concepts and measurement. New and completely revised edition.* Princeton, NJ: Princeton University Press.

Groestad, P. (2022). *Cognitive Warfare – Hacking the OODA Loop*. Estonia: NATO CCDCOE CyCon. Available at: https://www.youtube.com/watch?v=H3RqF5PiqXM (accessed October 27, 2023).

Hellström, J., Kallioniemi, P., Kytöneva, S., and Puranen, M. (2024). *StratCom | NATO Strategic Communications Centre of Excellence Riga, Latvia*. Riga, Latvia: NATO Strategic Communications Centre of Excellence. Available at: https://stratcomcoe.org/publications/are-russian-narratives-amplified-by-prc-media-a-case-study-on-narratives-related-to-swedens-and-finlands-nato-applications/298 (accessed April 2, 2024).

Hong, Z. (2013). The South China Sea dispute and China-Asean relations. *Asian Aff.* 44, 27–43. doi: 10.1080/03068374.2012.760785

Hung, T. C., and Hung, T. W. (2022). How China's cognitive warfare works: a frontline perspective of Taiwan's anti-disinformation wars. *J. Glob. Sec. Stud.* 7:ogac016. doi: 10.1093/jogss/ogac016

Jaitner, M. L., and Kantola, H. (2016). Applying principles of reflexive control in information and cyber operations. *J. Inf. Warfare* 15, 27–38. Available at: https://www.jstor.org/stable/26487549

Le Guyader, H. (2022). "Cognitive domain: a sixth domain of operations," in *Cognitive Warfare: The Future of Cognitive Dominance*, eds. B. Claverie, B. Prebot, N. Buchler, and F. du Cluzel (NATO Collaboration Support Office), 1–5. Available at: https://hal.science/hal-03635898v1

Lee, S. (2014). China's 'three warfares': origins, applications, and organizations. *J. Strat. Stud.* 37, 198–221. doi: 10.1080/01402390.2013.870071

Miller, S. (2023). Cognitive warfare: an ethical analysis. *Ethics Inf. Technol.* 25:46. doi: 10.1007/s10676-023-09717-7

Mueller, R. S. (2019). *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. Washington DC: U.S. Department of Justice, 448.

Muradov, I. (2022). The Russian hybrid warfare: the cases of Ukraine and Georgia. *Defence Stud.* 22, 168–191. doi: 10.1080/14702436.2022.2030714

NATO (1949). *The North Atlantic Treaty*. Available at: https://www.nato.int/cps/en/natohq/official_texts_17120.htm (accessed October 27, 2023).

NATO (2021). *Warfighting Capstone Concept*. Norfolk, VA: NATO ACT (2021). Available at: https://www.act.nato.int/our-work/nato-warfighting-capstone-concept/ (accessed August 23, 2023).

NATO (2022). *NATO 2022 Strategic Concept*. Available at: https://www.nato.int/nato_static_fl2014/assets/pdf/2022/6/pdf/290622-strategic-concept.pdf (accessed August 16, 2023).

NATO Allied Command Transformation (2021). *Concept Development and Experimentation Handbook*. Norfolk, VA: NATO Allied Command Transformation (ACT). Report No.: Version 2.10. Available at: https://www.act.nato.int/wp-content/uploads/2023/05/NATO-ACT-CDE-Handbook_A_Concept_Developers_Toolbox.pdf (accessed April 4, 2024).

NATO Allied Command Transformation (2023). *Cognitive Warfare Exploratory Concept. NATO Allied Command Transformation (ACT). Report No.: ACT/SPP/CNDV/TT-6700.*

Oksanen, P., Ålander, M., Eggen, K. A., Serritzlev, J., Kohv, M., Kjartansson, B. B., et al. (2024). *Tracking the Russian Hybrid Warfare – Cases From Nordic-Baltic Countries*. Stockholm: FRIVÄRLD, 39. Available at: https://frivarld.se/rapporter/tracking-the-russian-hybrid-warfare-cases-from-nordic-baltic-countries/ (accessed September 9, 2024).

Reding, D. F., and Wells, B. (2022). "Cognitive warfare: NATO, COVID-19 and the impact of emerging and disruptive technologies," in

COVID-19 Disinformation: A Multi-National, Whole of Society Perspective (Advanced Sciences and Technologies for Security Applications), eds. R. Gill, and R. Goolsby (Cham: Springer International Publishing), 25–45.

Sartori, G. (1970). Concept misformation in comparative politics. *Am. Polit. Sci. Rev*. 64, 1033–1053. doi: 10.2307/1958356

Sartori, G. (ed.). (1984). *Social Science Concepts: A Systematic Analysis*. Beverly Hills, CA: Sage, 455.

Splidsboel Hansen, F. (2021). When Russia wages war in the cognitive domain. *J. Slavic Milit. Stud*. 34, 181–201. doi: 10.1080/13518046.2021.1990562

Tashev, B., Purcell, M., and McLaughlin, B. (2019). Russia's information warfare: exploring the cognitive dimension. *Mar. Corps Univ. J*. 10, 129–147. doi: 10.21140/mcuj.2019100208

Wardle, C., and Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Strasbourg: Council of Europe. Available at: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html (accessed April 26, 2021).

Wonka, A. (2007). "Um was geht es? Konzeptspezifikation in der politikwissenschaftlichen Forschung," in *Forschungsdesign in der Politikwissenschaft: Probleme; Strategien; Anwendungen*, eds. T. Gschwend, and F. Schimmelfennig (Frankfurt am Main: Campus Verlag), 63–89.

frontiers | Frontiers in Artificial Intelligence

TYPE Methods
PUBLISHED 20 December 2024
DOI 10.3389/frai.2024.1509179

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Nikos Kanakaris,
University of Patras, Greece
George Tsihrintzis,
University of Piraeus, Greece

*CORRESPONDENCE
Loukas Ilias
✉ lilias@epu.ntua.gr

RECEIVED 10 October 2024
ACCEPTED 27 November 2024
PUBLISHED 20 December 2024

CITATION
Tzoumanekas G, Chatzianastasis M, Ilias L, Kiokes G, Psarras J and Askounis D (2024) A graph neural architecture search approach for identifying bots in social media.
Front. Artif. Intell. 7:1509179.
doi: 10.3389/frai.2024.1509179

A graph neural architecture search approach for identifying bots in social media

Georgios Tzoumanekas[1], Michail Chatzianastasis[2], Loukas Ilias[1]*, George Kiokes[3], John Psarras[1] and Dimitris Askounis[1]

[1]Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, [2]DaSciM, LIX, Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France, [3]Laboratory of Electrical Machines and Installations, Division of Electrical, Electronics and Informatics, School of Engineering, Merchant Marine Academy of Aspropyrgos, Aspropyrgos, Greece

Social media platforms, including X, Facebook, and Instagram, host millions of daily users, giving rise to bots automated programs disseminating misinformation and ideologies with tangible real-world consequences. While bot detection in platform X has been the area of many deep learning models with adequate results, most approaches neglect the graph structure of social media relationships and often rely on hand-engineered architectures. Our work introduces the implementation of a Neural Architecture Search (NAS) technique, namely Deep and Flexible Graph Neural Architecture Search (DFG-NAS), tailored to Relational Graph Convolutional Neural Networks (RGCNs) in the task of bot detection in platform X. Our model constructs a graph that incorporates both the user relationships and their metadata. Then, DFG-NAS is adapted to automatically search for the optimal configuration of Propagation and Transformation functions in the RGCNs. Our experiments are conducted on the TwiBot-20 dataset, constructing a graph with 229,580 nodes and 227,979 edges. We study the five architectures with the highest performance during the search and achieve an accuracy of 85.7%, surpassing state-of-the-art models. Our approach not only addresses the bot detection challenge but also advocates for the broader implementation of NAS models in neural network design automation.

KEYWORDS

bot detection, graph neural networks, neural architecture search, propagation, transformation, social media platform X

1 Introduction

Social media are online community platforms and apps that let users create, share, and interact with each other's content. Social media content can be text, photos, videos, GIFs, audio, etc. Social media can be used for various reasons, from users who share interests communicating to getting informed about current worldwide events. Social media can also be used for detecting early signs of stress and depression (Ilias and Askounis, 2023b; Ilias et al., 2024b; Kerasiotis et al., 2024). The existence of social media in our day-to-day lives is more prevalent than ever. As of 2023, there are roughly 4.9 billion social media users, a percentage that is more than 60% of the entire population and more than 100 social media platforms. X, previously known as Twitter, stands out as one of the most widely recognized social media platforms. Twitter was launched in 2006. It revolves around the concept of "following" other users. A user can follow accounts they are interested in and

see their tweets in their timeline ("following") and conversely can have "followers" that see their tweets. Nowadays, it has been established as a powerful tool for real-time news updates, public discourse, and social movements, and continues to evolve and enhance its user experience. In 2023, Twitter was renamed to X by then-CEO Elon Musk. The extensive presence of social media in the modern landscape has led to the emergence of accounts that automate interactions on social media platforms, often mimicking human behavior, the so-called bots. These bots can be coded to perform a variety of tasks, such as automatically publishing content, liking, sharing, following, or commenting on posts. Some can even be programmed to engage in conversations to promote specific agendas. Their behavior differs depending on their intent and purpose, but they might share features, such as very high or very low activity levels and more structured and characteristic language patterns (Alsmadi and O'Brien, 2020). Uyheng et al. (2022) examined the origin and traits of trolling messages, finding that they often originate from automated bots and are distinguished by their use of abusive language, reduced cognitive complexity, and specific targeting of individuals or entities. Their study also noted a tendency for bots to target right-leaning sources of information, while trolls tended to engage with less polarized content, spreading misinformation across diverse audiences. Bots are very efficient in spreading misinformation, particularly when programmed with optimized values for factors like walking speed, network distribution, and strategy (Zhang et al., 2024). Fake news and bots have had significant tangible consequences in several cases. Users tend to believe conspiracy theories and misinformation, and correction attempts can sometimes backfire (Xu et al., 2023). Users might also share fake news for altruistic or self-promotional purposes, yet those with greater social media literacy are better equipped to identify and refrain from spreading fake news (Mi and Apuke, 2024). Therefore, there's a need for measures to promote truthful reporting in media and detect any cases of misinformation dissemination.

The need to detect bot accounts to shut them down is quite immediate, assessing the hazards of their uncontrollable presence on social media. Several studies to identify bots from real users have been conducted that provide satisfactory results. There have been several approaches, including supervised learning (Lee et al., 2021), unsupervised learning (Cresci et al., 2017), reinforcement learning (Alauthman et al., 2020), and GNN-based architectures (Feng et al., 2021c). However, all these traditional neural architectures often rely on fixed parameters that are manually designed. Constructing efficient neural network architectures requires extensive feature engineering and can be a quite challenging and time-consuming procedure. Also, fixed architectures often mitigate the models' adaptability on other datasets and tasks. Motivated by these limitations, we examine the implementation of Neural Architecture Search (NAS) to automate the process of discovering optimal architectures. NAS explores a search space of possible architectures and identifies the configurations that enhance the model's performance.

A Neural Architecture Search method that has been proposed to solve the performance issues of fixed architectures is Deep and Flexible Graph Neural Architecture Search (DFG-NAS) (Zhang et al., 2022). It employs an evolutionary algorithm to explore a vast space of permutations of Propagation and Transformation operations, to find the one with the best accuracy in the validation set. Addressing the limitations of previous bot detection models due to their fixed architectures we employ DFG-NAS on a GNN-based approach for bot detection. This approach leverages the user's semantical and property information and constructs a heterogeneous graph out of the follower-following relationships between users. Then, we adapt the DFG-NAS approach to handle Relational Graph Convolutional Neural Networks (RGCNs). The model automatically searches for the permutation of Propagation (P) and Transformation (T) functions, the two main processes of the message-passing protocol, with the highest validation accuracy. The model is also amplified with the use of the Gate operation on the P connections and the use of the skip-connection operation on the T connections.

To the authors' knowledge, DFG-NAS has not been employed before in the task of bot detection. All our experiments were performed on the Twibot-20 dataset (Feng et al., 2021b). The following sums up the contributions of our work:

- We implement DFG-NAS, tailored to RGCNs, to automatically determine the most effective permutation of the message-passing operations.
- We perform experiments to demonstrate the benefits of architecture search in bot detection and compare our method to state-of-the-art models.
- We perform a thorough ablation study on the necessity of the user metadata in our graph, the Gate operation, and the skip-connection operation in NAS.

2 Research objective

It is evident to any social media user that bots continue to dominate the digital landscape despite extensive efforts in bot detection and platform initiatives to stop their activities. As technology advances, bots are programmed to mimic human mannerisms more effectively, making them more resilient against detection mechanisms. Beyond the irritation they pose to everyday users, some bots can have tangible and detrimental effects on human society. In 2016 fake news stories spread widely during the U.S. presidential election campaign aiming to influence the public vote (Bessi and Ferrara, 2016). Throughout the COVID-19 pandemic (Ferrara, 2020), bots spread misinformation about the virus and the vaccines on social media, leading to mob panic, confusion, and even resistance to public health measures. Fake news is often framed in a manner that fosters negativity in social discussions and hinders individuals' ability to consider diverse perspectives, contributing to the formation of "echo chambers" on social media platforms (Scheibenzuber et al., 2023). Bots also exacerbate cyberbullying by mass-targeting users, leading to serious psychological consequences. Social media platforms face challenges in effectively moderating such content. Cyberbullying detection methods often rely on unclear definitions and are prone to biases in data annotation (Mahmud et al., 2023). Their evolving nature raises concerns about the efficiency of current preventive measures,

highlighting the need for innovation to prevent the dangers posed by this digital phenomenon.

The motivation for this research was constructing a model characterized by adaptability across future datasets, ensuring resilience in the face of evolving technology through time. Many contemporary models rely on fixed architectures, often struggling to demonstrate their efficiency on novel datasets. Although Neural Architecture Search (NAS) has shown significant advantages in various test cases, its application to bot detection remains relatively underexplored, with limited but promising results noted in studies such as Yang et al. (2023). Considering the dynamic nature of the social media landscape and the continuous evolution of bots, more flexible architectures specifically designed for bot identification could offer a practical solution to mitigate their real-world consequences.

This research aims to showcase the efficiency advantages of architecture search and perhaps pave the way for more implementations of NAS models in bot detection in the ongoing battle against automated malicious activities.

3 Related work

3.1 Bot and fake news detection models

The task of bot identification has attracted numerous studies and many state-of-the-art models propose fascinating methodologies. We could mainly divide these models into supervised learning approaches, unsupervised learning approaches, and GNN-based approaches. In this section, we present some baseline models proposed for bot detection and discuss how they fall into the above categories.

Lee et al. (2021) applied various machine learning algorithms, including SVMs, Naive Bayes, and decision trees, to build and evaluate a supervised bot detection model. The features used in their analysis included account-based features (e.g., the number of followers, friends, tweets), temporal features (e.g., time of account creation, tweet frequency), and content-based features (e.g., usage of URLs, hashtags). Kudugunta and Ferrara (2018) suggested a deep learning model that uses the user's tweets and some metadata features. This architecture includes a tokenizer, GloVE embedding layer, LSTM, and Dense layers. Wei and Nguyen (2019) used only users' tweets with no context of prior knowledge on user profiles, friendship networks, or behavior. They proposed a recurrent neural network (RNN) model that used word embeddings to encode tweets, a three-layer Bidirectional LSTM (BiLSTM), and a softmax layer at the binary output. Cai et al. (2021) proposed their model (BeDM) that involved deep neural networks in bot detection. They employed convolutional neural networks (CNNs) and LSTM, using only the tweet semantics, such as the frequency and the type of publications. Botometer is a web-based program developed by Davis et al. (2016) at Indiana University. It leverages more than 1,000 features to classify Twitter accounts as bots and humans, such as friends, the structure of the social network, user meta-data, temporal activity, and sentiment analysis. Botometer distinguishes the accounts by an overall bot score (ranging from 0 to 5), along with several other scores. The greater the score, the greater the probability that this account is linked to a bot. Yang et al. (2022) presented a thorough introduction of the latest version of Botometer for new users and demonstrated a case study. Alarfaj et al. (2023) utilized features based on content attained from the Twitter API and employed state-of-the-art classifiers, like MLPs, random forest, and naive Bayes. Features included messages, special characters, sentiment analysis, etc. Alothali et al. (2022) introduced their framework, called Bot-MGAT, which stands for bot multi-view graph attention network. The scientists pointed out that other approaches couldn't adjust to different datasets since there wasn't enough recently updated labeled data, which made sense given the constantly shifting behavior of the bots. They presented a methodology that makes use of transfer learning (TL) to leverage the multi-view graph attention mechanism. The framework also benefited from semi-supervised learning, using labeled and unlabeled data. The authors used the TwiBot-20 (Feng et al., 2021b) due to its graph structure, extracting 18 features for the training. Feng et al. (2021a) suggested SATAR. In particular, SATAR leverages the user's semantics, property, and neighborhood information. It adjusts by fine-tuning parameters and pre-training on a huge number of self-supervised users. The authors proposed two alternative models: $SATAR_{FC}$ and $SATAR_{FT}$. Ilias and Roussaki (2021) proposed two methods for bot detection using deep learning techniques. Their first approach extracts a substantial 71 features per user to utilize for account classification to bots and genuine users. They also employed various feature selection techniques to discard redundant and irrelevant features. Their second methodology proposes a deep learning architecture for tweet-level classification. This architecture incorporates an attention mechanism atop the Bidirectional Long Short-Term Memory (BiLSTM) layer. During the learning phase, the attention mechanism helps the model better focus on relevant information. Ilias et al. (2024a) focused solely on user descriptions and sequences of actions performed by Twitter accounts. Their approach includes both unimodal (text or image) and multimodal (both text and image) methods. They designed digital DNA sequences per user based on tweet type and content, converted these sequences into 3D images, and fine-tuned pre-trained vision models like AlexNet, ResNet, and VGG16. For bot detection through user descriptions, they fine-tuned TwHIN-BERT, a transformer model. In multimodal approaches, they use VGG16 for visual representation and TwHIN-BERT for textual representation, proposing three fusion methods: concatenation, gated multimodal unit (GMU), and cross-attention. They conducted their experiments on both the Cresci'17 and TwiBot-20 dataset. Wei and Nguyen (2023) proposed their model BOTLE. Their model utilizes a recurrent neural network (RNN) with Bidirectional Gated Recurrent Units (BiLGRU) connecting two hidden layers of opposite directions leading to the same output. Notably, BOTLE does not rely on handcrafted features or pre-existing information regarding account profiles. Linguistic embeddings, including word, character, part-of-speech, and named-entity embeddings, are employed to encode tweet content, eliminating the need for labor-intensive feature engineering. Bazmi et al. (2023) introduced the Multi-View Co-Attention Network (MVCAN), which aims to capture the latent

topic-specific credibility of both users and news sources. This model represents news articles, users, and news sources in a manner that encodes topical viewpoints, socio-cognitive biases, and partisan biases as vectors. These features are encoded using a variant of the Multi-Head Co-Attention (MHCA) mechanism. Shevtsov et al. (2023) introduced their model BotArtist, constructed on a semi-automatic machine learning pipeline, that requires minimal features for training, taking into consideration the loads of data needed by previous approaches and the recent monetization of Twitter API requests. Sujith et al. (2022) proposed a supervised learning approach that used multiple models to detect bots. Their classification of accounts relied on features like user metadata, tweet content, and posting history, among others. In addition to identifying bot accounts, the authors assigned a level of significance or influence to them, prioritizing the removal of the most influential or harmful bot accounts. Liu et al. (2023) proposed BotMoE, which leverages three perspectives of user information (metadata, text, and graph representations) and incorporates a community-aware Mixture-of-Experts (MoE) layer to assign users to different communities. The user representations are fused with an extractor fusion layer and supervised learning is employed to train the BotMoE framework to perform community-aware bot detection. Saxena et al. (2023) proposed two frameworks for recognizing accounts that disseminate false information on Twitter. Initially, they employed profile-based data, including the verified status, profile photo, and account lifetime and activity. Then, they combined tweet-propagation patterns and assigned a credibility score to each user, signifying their authenticity. Dimitriadis et al. (2024) proposed CALEB that is based on the Conditional Generative Adversarial Network (CGAN) and its extension, Auxiliary Classifier GAN (AC-GAN). By developing realistic artificial bot varieties, they were able to replicate the evolution of bots. As a result, they enhanced already-existing datasets and were able to identify bots before they emerged.

Yang et al. (2020) used a combination of unsupervised and supervised learning methods for bot detection. Specifically, the authors utilized minimal features derived from user metadata, temporal patterns, network structure, sentiment analysis, and linguistic cues that they fed into a machine learning pipeline, that reduced dimensionality and included classification algorithms. Cresci et al. (2017) introduced the Social Fingerprinting technique for bot detection, a Digital DNA technique that models social network users' behaviors. Each user is represented as a sequence of characters depending on the type and content of the tweets they publish, simulating a DNA sequence. The authors try to find similarities in the sequences defining the length of the Longest Common Substring (LCS) between two sequences. For a set of real users, the length of LCS was found to be particularly small, leading to the conclusion that longer sequences than the average LCS were bots. Based on this idea, the authors developed two techniques, one based on supervised learning and another on unsupervised learning to find similarities in the behavior of accounts. Quezada et al. (2023) developed a real-time bot infection detection model that analyzes Domain Name System (DNS) traffic events. They extracted 13 attributes from DNS logs to create unique fingerprints for servers. Using Isolation Forest, an algorithm for unsupervised learning, they identified anomalies in the fingerprints to classify hosts as infected or not. The model also utilized Domain Generation Algorithms (DGA) to search for queries to anomalous domains. Finally, a Random Forest, a supervised learning algorithm, was employed to create a model for detecting future bot infections on hosts. Miller et al. (2014) approached bot identification as an anomaly detection problem. They extracted 107 features from user's tweets and property information and adapted two stream clustering algorithms, StreamKM++ and DenStream, to facilitate spam detection and identified bot users as abnormal outliers. Chavoshi et al. (2016) developed DeBot, a bot detection system for social media, using warped correlation to identify likely bot accounts based on their high synchronicity, a characteristic unlikely in human users. DeBot doesn't require labeled data and operates on activity correlation. Moreover, through the utilization of a lag-sensitive hashing technique, it can promptly cluster accounts for real-time classification. Minnich et al. (2017) proposed their real-time unsupervised model BotWalk. Using metadata, content, temporal, and network features they employ anomaly detection, comparing each user to a seed bank of labeled accounts iteratively. Mannocci et al. (2022) proposed MulBot, an unsupervised bot detection system that utilizes multivariate time series (MTS) analysis. They employed an LSTM autoencoder to map the MTS data into a latent space and then conducted clustering on this encoded data to find dense clusters of users exhibiting similar behavior, assuming this was a common trait of bot accounts. MulBot also showcases effectiveness in identifying and distinguishing various botnets. Wu et al. (2022) employed unsupervised machine learning techniques, specifically K-Means and Agglomerative clustering, for Twitter bot detection. They used account activity, popularity, and verification status, among other features for the clustering. Koggalahewa et al. (2022) introduced an unsupervised method for bot identification based on a user's peer approval in the social network. They based peer acceptance between two users on their shared interests over a multitude of issues. Lopes et al. (2022) introduced their botnet identification model, designed to detect networks of compromised devices under master control. Their approach relies on analyzing network flow behavior through a contemporary method known as the Energy-based Flow Classifier (EFC). EFC employs inverse statistics to enhance anomaly detection.

Ali Alhosseini et al. (2019) introduced the use of graph convolutional neural networks (GCNN) in bot identification. They noted that besides the users' features, the construction of a social network would enhance a model's ability to distinguish the bots from the genuine users. Feng et al. (2022) introduced the aspect of diversity in relationships and influence dynamics among users in the Twittersphere for bot detection. They proposed a bot detection framework that leverages a network with users as nodes and the different relations as edges. Then they aggregated messages across users and operated heterogeneity-aware Twitter bot detection. They conducted their experiments using the Twi-Bot20 dataset. Feng et al. (2021c) proposed their model for bot detection BotRGCN, which is short for Bot detection with Relational Graph Convolutional Networks. BotRGCN builds a heterogeneous graph out of the following relationships and uses information, such as the user's description, tweets, numerical

and categorical property set, and neighborhood information. The experiments were conducted on the Twi-Bot20 dataset (Feng et al., 2021b), but BotRGCN could exploit other types of relations if supported by the dataset. Kušen and Strembeck (2020) examined the structural dynamics of conversations between humans and bots on Twitter following emotionally charged riot events. They introduced "emotion-exchange motifs" to identify recurring patterns in emotional message exchanges. Their findings revealed that human conversations exhibited various motifs with reciprocal edges and self-loops, indicating interactive dialogue. In contrast, bots typically disseminated identical messages to multiple users or did not anticipate replies. Moreover, bots frequently initiated conversations and often conveyed fear-inducing messages. Bui and Potika (2022) introduced a graph-based method for bot detection. They detailed their data collection process and identified specific behaviors indicative of an account being associated with a bot. These behaviors can include engagement with other users, nonsensical usernames and profile information, repetitive content posting, and retweeting activity. These observations are utilized to label the accounts accordingly. Dehghan et al. (2023) suggested that the local social network formed around each account can aid in identifying the bots. To prove their hypothesis, they compared two classes of embedding algorithms, the former of which focused on proximity data and the latter that focused on nodes' neighborhoods. They discovered that the structural embeddings presented higher information underlining the valuable information that is embedded within each node's local network. Pham et al. (2022) introduced their approach Bot2Vec, which eliminated the need for user profile features. To improve the model's generalization on many social media platforms, they used only local neighborhood relations and the community structure of the graph that represented the users and employed an random walk strategy in the communities. Noekhah et al. (2020) proposed their model "Multi-iterative Graph-based opinion Spam Detection" (MGSD) that aims to identify various types of spam entities. It analyzes all kinds of relationships between them and utilizes domain-independent features, allowing for generalization across types of opinionated documents. Trained on both existing and novel features, MGSD assigns a spam score to each entity. Ye et al. (2023) proposed HOFA, a graph-based framework for bot detection, featuring two key modules: Homophily-Oriented Graph Augmentation (Homo-Aug) and Frequency Adaptive Attention (FaAt). The Homo-Aug employs an MLP to extract user representations and generate a k-NN graph. Meanwhile, the FaAt module acts as a low-pass filter for homophilic edges and a high-pass filter for heterophilic edges. This function prevents excessive smoothing of user features by the neighborhood. El-Mawass et al. (2020) explored using the output of existing supervised classification systems to detect spammers. They incorporated the classifiers' outputs as prior beliefs within a probabilistic graphical model framework. Proposing a bipartite users-content interaction graph, they facilitated the spread of beliefs to similar accounts. Constructing a Markov Random Field on a graph of similar users, they employed Loopy Belief Propagation to derive the predictions. Their findings demonstrated a notable enhancement in recall while maintaining precision.

3.2 Neural architecture search approaches

Neural architecture search can increase performance in many tasks (Chatzianastasis et al., 2023). Graph neural architecture search is proposed as the solution to performance limitations due to a fixed architecture. Parameter tuning in neural networks can be a challenging task. Many NAS methods have been suggested that include variations in the search space, the optimization method, and the architecture evaluation. We will divide these methods based on their optimization method, which will include reinforcement learning, evolutionary algorithms and gradient-based methods.

Zhou et al. (2022) proposed the automated graph neural networks (Auto-GNN) framework. Auto-GNN searches for the best GNN architecture possible in a predetermined search space, divided into six classes of actions: hidden dimension, attention function, attention head, aggregate function, combine function, and activation function. For efficiency reasons, the authors designed a conservative explorer to preserve the optimal neural architecture discovered during the search. The authors also implemented constrained parameter sharing, adapted to the heterogeneous GNN architecture. Two experimental methods were presented: inductive, in which the graph structure and node features on the testing and validation sets are unknown during training, and transductive, which involves the availability of unlabeled data for testing and validation during training. Gao et al. (2021) proposed GraphNAS to implement an automatic search of the best graph neural architecture based on reinforcement learning. The search space covers sampling functions, aggregation functions, and gated functions. GraphNas also uses more efficient parameter-sharing techniques than other contiguous models for CNNs and RNNs. After training 1,000 different architectures, the five best ones were used for the testing, which surpassed human-invented ones or those produced by random searches. Zhao et al. (2020) proposed the SNAG framework (Simplified Neural Architecture Search for Graph neural networks). The suggested framework had two key components: Node aggregators, which focused on neighborhood features, and Layer aggregators, which focused on the range of the neighborhood used. The search space algorithm was a variant of Reinforcement Learning that adopted the weight-sharing mechanism (SNAGWS). Nunes and Pappa (2020) presented one NAS methods for optimizing GNNs based on reinforcement learning and one based on evolutionary algorithms. The authors defined two cases of search spaces: Macro, where the architectures generated have the same structure, and Micro, where the architectures are not rigidly structured but combine several convolutional schemas. They concluded that EA and RL found very similar architectures to those found by a random search, a significantly simpler technique. However, they pointed out that whilst the other approaches generated large structures in as much as 80% of the situations, EA created the majority of GPU-fitting designs. Li et al. (2023) proposed Meta-GNAS that uses meta-reinforcement learning from past tasks to apply that knowledge to new tasks. Additionally, they speed up the search by using a predictive model to evaluate the potential graph neural architectures instead of training them from scratch.

Peng et al. (2020) implemented a NAS approach to human action recognition from skeleton movements. The search space

was enlarged with diverse spatial-temporal graph modules while constructing higher-order connections between nodes using Chebyshev polynomial approximation. The search algorithm used is an evolutionary adaptation with a high sampling efficiency, denoted Cross-Entropy method with ImportanceMixing (CEIM). Jiang and Balaprakash (2020) adapted the method of neural architecture search to the conception of GNNs for predicting molecular properties. The authors designed neural networks for message-passing (MPNNs) between nodes. To find an optimal MPNN from the user-defined search space, they used regularized evolution (RE) from the DeepHyper package. Zhang et al. (2022) proposed DFG-NAS, an innovative method that allows for automatic search of very deep and adaptable GNN architectures. DFG-NAS focuses on exploring macro-architectures, specifically the implementation details of atomic propagation (P) and transformation (T) operations within the GNN. P is linked to the graph structure, whereas T concentrates on the non-linear transformations within the neural network. In addition, they adopted gating and skip-connection mechanisms for deeper GNN pipelines. They used an evolutionary algorithm to find the optimal architecture, which supported four cases of mutation. Peng et al. (2023) introduced Fast-ENAS as a computationally efficient alternative to Evolutionary Neural Architecture Search. This method utilizes a training-free performance metric that is computed with a single forward pass. The authors enhance the search process by incorporating a GCN-based contrastive predictor, aiming to improve the accuracy of the estimated performance of a candidate architecture, bringing it closer to its actual performance. Shang et al. (2023) introduced AG-ENAS, which brings two key innovations to the Evolutionary Neural Architecture Search process. Firstly, it employs an adaptive parameter adjustment mechanism based on population diversity and fitness, enhancing the adaptation of genetic operators' associated parameters. Secondly, the model introduces a mutation operator guided by the gene potential contribution. It improves offspring quality by assigning weight to more valuable genes through a distribution index matrix. The concept of aging is integrated into environmental selection to mitigate premature convergence. Lopes et al. (2024) presented Guided Evolutionary Architecture (GEA), which tackles the problem of other NAS models getting trapped in suboptimal solutions during the search process. GEA overcomes this challenge by generating and evaluating multiple architectures using a zero-proxy estimator and selecting only one with the best-performing one for the next generation. This approach expands the search space without increasing complexity, as new architectures are derived from previous ones through mutations.

Zhao H. et al. (2021) proposed their framework SANE. The search space has similarities with the search space from the SNAG framework, with Node and Layer aggregators. However, the authors presented a novel differentiable search algorithm. Cai et al. (2021) introduced a GNAS approach featuring a uniquely designed search space and a gradient-based search approach. The authors developed a three-level Graph Neural Architecture Paradigm (GAP) that includes two types of fine-grained atomic operations (neighbor aggregation and feature filtering) that are derived from message-passing, to build the search space. Li et al. (2021) introduced an innovative dynamic one-shot search space designed for multi-branch neural architectures within GNNs. The dynamic nature of the search space offers a larger capacity than a larger predefined search space. The architectures with lower importance weights are removed periodically from the population, while the candidate operations are unique to every edge of the computational graph. The authors performed both supervised and unsupervised techniques for the training part. Zhao J. et al. (2021) proposed a gradient-based architecture search method for predicting a system's remaining useful life. Their approach models the search space as a directed acyclic graph (DAG), where nodes represent latent representations and edges represent transformation operations. By employing candidate operations like ReLU and tanh, along with the softmax function, they make the search space continuous and the objective function differentiable, facilitating gradient-based optimization methods to find the optimal architecture.

3.3 Related work review findings

From the aforementioned research works, it is clear that there have been many approaches to the task of bot detection. Previous studies include supervised, unsupervised, and graph neural network (GNN) based methods. While they have shown promising results, the relentless evolution of bot accounts toward simulating human-like patterns poses a significant challenge to their effectiveness. These models are constrained by fixed architectures, limiting their adaptability to newer datasets.

Little work has been done in employing Neural Architecture Search methods in GNN-based methodologies for bot detection. Our work shifts the focus on overcoming the performance limitations due to fixed architectures, by utilizing DFG-NAS to search for the best configuration of Propagation and Transformation functions in the message passing protocol of our RGCNs. Instead of extensive feature engineering our model searches for the permutation with the highest accuracy and aims for better adaptability in newer datasets that will depict future bots' behavior. Moreover, DFG-NAS presents high advantages, as it is suitable for GNN-based methods and overcomes over-smoothing and model degradation issues with the gate and skip-connection operations.

4 Dataset

The TwiBot-20 Dataset (Feng et al., 2021b) is a publicly available dataset, constructed with a breadth-first search (BFS) methodology. The dataset includes information about each user's profile information obtained from the Twitter API, recent tweets, and domains of the user's interest. It also contains information about the user's neighborhood, which helps us construct a heterogeneous graph from the following relationships. Table 1 presents all the attributes of the TwiBot-20 Dataset and a short description of them. The information from the user profiles is further mentioned in the preprocessing part of the model. The graph that is constructed consists of 229,580 nodes and 227,979 edges. The objective of the bot detection system is to distinguish between bots and genuine users by analyzing information from user

TABLE 1 TwiBot-20 dataset attributes.

Attribute	Description
ID	ID from Twitter to identify the user
Profile	Profile information from Twitter API
Tweet	200 recent tweets of the user
Neighbor	20 random followers and followings of the user
Domain	Domain of the user (politics, business, entertainment, sports)
Label	Label of the user ("1": bot, "0":human)

descriptions, tweets, numerical and categorical properties, as well as neighborhood information.

5 Methodology

In this part, we present a complete analysis of our methodology. First, we describe the preprocessing of the user metadata used in our model. Next, we introduce the use of Relational Graph Convolutional Neural Networks and the two functions in Message Passing. Last, we explain the use of DFG-NAS (Zhang et al., 2022) in searching for the best permutation of Propagation and Transformation functions. In Figure 1, we depict the architecture of the model on a higher level, while Figure 2 presents the connections between the different layers of an example configuration of P and T functions.

5.1 Data preprocessing

We follow the preprocessing suggested by Feng et al. (2021c) for BotRGCN. Each user's representation includes metadata that are preprocessed as follows:

- Overall: user's description, tweets, numerical and categorical properties are encoded and concatenated to finally represent the user's metadata:

$$r = [r_b; r_t; r_p^{num}; r_p^{cat}] \in \mathbb{R}^{D\times 1} \tag{1}$$

  where D is the user embedding dimension. Each feature's procession and representation are explained below. Later we will prove that the model's performance is attributed to all these features and not only to the heterogeneous graph.
- User description: the user descriptions are encoded with pre-trained RoBERTa:

$$\bar{b} = RoBERTa(\{b_i\}_{i=1}^{L}), \bar{b} \in \mathbb{R}^{D_s\times 1} \tag{2}$$

  where $\bar{b}$ denotes the user description representation and D_s is the dimension of the RoBERTa embedding. The vectors for the user's description are derived:

$$r_b = \phi(W_B \cdot \bar{b} + b_B), r_b \in \mathbb{R}^{D/4\times 1} \tag{3}$$

  where W_B and b_B represent trainable parameters, ϕ denotes the activation function, and D is the dimension of the embedding.
- User tweets: the user tweets are also encoded using RoBERTa. The ultimate representation of the user's tweets, denoted as r_t, is computed as the average of the representations of all individual tweets.
- User numerical properties: the user's numerical properties are adopted straight from the Twitter API with no feature engineering and presented in Table 2. For this information z-score normalization is conducted to get the representation r_p^{num} from a fully connected layer.
- User categorical properties: the user's categorical properties are also encoded with MLPs and GNNs, without feature engineering, just as the numerical properties. They are adopted straight from the Twitter API and presented in Table 3. After one-hot encoding, they are concatenated and transformed through a fully connected layer and leaky-relu to get their representation r_p^{cat}.

5.2 Relational graph convolutional neural networks

Our method builds a heterogeneous graph out of the following relationships. Users are considered nodes and the "following" and "followers" relations are represented as edges connecting the nodes. The user's "followers" are therefore represented differently than the user's "following." The heterogeneous graph that is constructed can represent better the relations between users and more relations between the users could be integrated into the graph if supported by the dataset. The users also contain the concatenated metadata that we described below.

To combine the users' representations with the relationships between users we make use of RGCNs. The message-passing process in RGCNs comprises two fundamental operations: propagation (P) of the representations of the user's neighbors and transformation (T) on these representations. Below we describe the process behind the two functions:

- Propagation (P): propagation includes message aggregation from neighbor nodes without explicit node feature transformation. The mathematical expression for the propagation step is as follows:

$$h_i^{(l+1)} = \sum_{r\in R}\sum_{j\in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} \tag{4}$$

  where $h_i^{(l+1)}$ is the new node feature after propagation, R is the set of relations, N_i^r are the neighbors of the node with relation r, $c_{i,r}$ is a normalization constant that can be learned or chosen in advance (for example $c_{i,r} = N_i^r$) and $W_r^{(l)}$ is the learnable weight matrix for relation r.
- Transformation (T): transformation occurs on each node based on the relations. The mathematical expression for the transformation step is as follows:

$$h_i^{(l+1)} = W_{root} h_i^{(l)} + \sum_{r\in R}(W_r h_i^{(l)}) \tag{5}$$

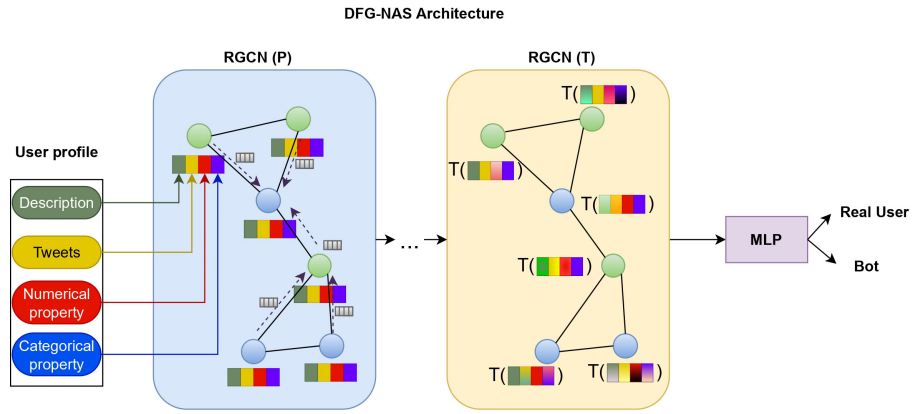

FIGURE 1A
Model used for Bot detection. User metadata is fed to the architecture proposed by NAS. The P step includes message aggregation from neighbor nodes. The T step includes the transformation process on each node based on neighbor relations. In the final part, an MLP decides whether the account belongs to a real user or a bot.

where $h_i^{(l+1)}$ is the new node feature after transformation, W_{root} is the learnable weight matrix for the root node, W_r is the learnable weight matrix for relation r and R is the set of relations.

We segregate these two types of functions since combinations of them will construct the search space for the architecture search.

5.3 Graph neural architecture search

The use of Graph Neural Networks offers undeniable advantages in the task of bot detection. However, maximizing their performance may require extensive feature engineering. This is why we employ Graph Neural Architecture Search, using the model DFG-NAS (Zhang et al., 2022). Thus, we search for the permutation of Propagation and Transformation steps that achieves the highest accuracy. Most G-NAS methods have a fixed pipeline length since the performance decreases with too many P operations as the layers become deeper, which is referred to as the over-smoothing issue. Propagation and transformation operations regulate the effect of smoothing. Moreover, with unlimited pipeline length DFG-NAS searches for more flexible pipelines of P and T operations, using an evolutionary algorithm. It also makes use of gating and skip-connection mechanisms in the P and T operations, respectively.

The search space includes P-T combinations and the number of P-T operations. The output of node v in the l-th layer is represented by $o_v^{(l)}$ in a single P or T operation within a single GNN layer of the model. The layer indices of all P and T operations are included in two sets, L_P and L_T. The connections of P and T are depicted in Figure 1B and also described below.

5.3.1 Propagation connections

An imminent problem in GNNs is over-smoothing or under-smoothing, a problem that arises with too many or too few propagation operations. To achieve suitable smoothness for different nodes, the P operations are amplified with a gating mechanism. If the next operation is also P, the result of the l-th P operation is the propagated node embedding of $o^{(l-1)}$. On the other hand, if T is the next operation, a node-adaptive combination weight is allocated for the node embeddings propagated by all of the previous P operations. Formulatively:

$$z_v^{(l)} = P(o_v^{(l-1)}) \tag{6}$$

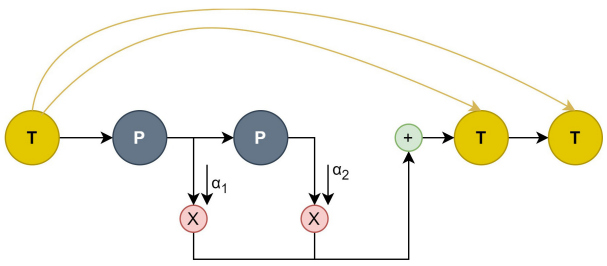

FIGURE 1B
Example of connections between the layers of NAS architecture. New T steps congregate information from all previous T steps. P steps propagate their embeddings and sum them up for the next T step.

TABLE 2 User numerical properties.

Feature name	Description
#followers	Number of followers
#followings	Number of followings
#favorites	Number of likes
#statuses	Number of statuses
active_days	Number of active days
screen_name_length	Screen name character count

TABLE 3 User categorical properties.

Feature name	Description
protected	Protected or not
geo_enabled	Geo-location enabled or not
verified	Verified or not
contributors_enabled	Enable contributors or not
is_translator	Is translator or not
is_translation_enabled	Translation or not
profile_background_tile	The background tile
profile_user_background_image	Background image or not
has_extended_profile	Extended profile or not
default_profile	The default profile
default_profile_image	The default profile image

$$o_v^{(l)} = \begin{cases} z_v^{(l)}, & followed\ by\ P \\ \sum_{i \in L_P, i \leq l} softmax(a_i) z_v^{(i)}, & followed\ by\ T \end{cases} \quad (7)$$

where $a_i = \sigma(s \cdot o_v^i)$ represents the weight for the i-th layer output of node v. Here, s is the learnable vector shared among the entirety of nodes, and σ denotes the Sigmoid function. To ensure proper scaling, the Softmax function is employed to normalize the sum of gating scores, making it equal to 1.

5.3.2 Transformation connections

An imminent issue with GNNs is the model degradation issue, caused by a hyperbolic amount of transformation operations and may result in a reduction of the model's accuracy. To mitigate this issue, skip-connection mechanisms are used in T operations. Each T operation's input is the total of all the T operations' outputs up to the last layer and the output from the layer before it. The input and output of the l-th T operation can be formulated as:

$$z_v^{(l)} = o_v^{(l-1)} + \sum_{i \in L_T, i < m(l)} o_v^{(i)} \quad (8)$$

$$o_v^{(l)} = \sigma(z_v^{(l)} w^{(l)}) \quad (9)$$

where $m(l)$ represents the index of the last T operation before the l-th layer, and $W(l)$ denotes the trainable parameter in the l-th T operation.

Evolutionary algorithms are a class of optimization algorithms inspired by biological evolution that aim to achieve the best accuracy in offspring through mutations. In our case, each GNN architecture is represented as a sequence of P and T operations. Each pipeline can be considered a chromosome and the mutations that occur simulate nature's mutations. These mutations can happen at any random position in the sequence. In our instance, four different cases of mutation can be enforced:

- +P: append a propagation operation.
- +T: append a transformation operation.
- P→T: replace a propagation operation with a transformation one.
- T→P: replace a transformation operation with a propagation one.

Initially, k distinct GNN designs are generated at random and evaluated on the validation set. These architectures represent the initial population set Q. Subsequently, m ($m < k$) members of the population are randomly sampled, and parent A is determined by selecting the member with the highest validation accuracy. By enforcing a random mutation of the four presented on A, a child architecture B is produced. B is then evaluated and added to the population, and the oldest person is eliminated. After T generations of this procedure, the architecture with the best performance is eventually returned.

DFG-NAS returns a sequence of P and T steps. As illustrated in Figure 1A, each step consists of an RGCN that conducts one of the two main functions as we described incorporating both the user metadata and the user relations. After the RGCNs layers an MLP is employed to finally distinguish bots from genuine users.

6 Experiments

6.1 Baselines

We compare our proposed apporach to the state-of-the-art models that are referenced in the paper of BotRGCN (Feng et al., 2021c). These experiments are all ran on the same dataset as the one we used for a fair comparison. We are using the published results for the comparison. More specifically, we compare our model to these state-of-the-art models:

- Lee et al. (2021) employed different supervised algorithms with several user features.
- Yang et al. (2020) used a combination of supervised and unsupervised learning with minimal user features.
- Kudugunta and Ferrara (2018) used both the tweets and the account metadata.
- Wei and Nguyen (2019) employed an RNN model utilizing only the user's tweets.
- Miller et al. (2014) extracted 107 features and employed stream clustering algorithms.
- Cresci et al. (2017) identified bots by computing the longest common substring between encoded sequences of users.
- Botometer (Davis et al., 2016) is a web-based program that leverages more than 1,000 user features.
- Ali Alhosseini et al. (2019) introduced graph convolutional neural networks in bot detection.
- SATAR (Feng et al., 2021a) leverages the user's semantics, property, and neighborhood information
- Feng et al. (2021c) used the user's description, tweets, numerical and categorical properties, and neighborhood information.
- Ilias et al. (2024a) designed two cross-attention layers based on the digital DNA sequence.

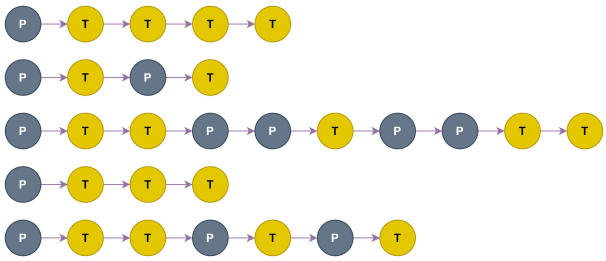

FIGURE 2
Permutations of Propagation (P) and Transformation (T) functions of the top-5 performing architectures from DFG-NAS. Their validation accuracies in the architecture search (from up to down) are: 87.01%, 86.99%, 86.95%, 86.89%, 86.82%.

6.2 Experiment settings

The experiment was run on Google Colab using Nvidia's T4 GPUs. The population set k for the architectural search is 15, and the maximum generation time T is 80. The training budget of each GNN architecture is 70 epochs. These numbers although limited due to our resources, provide a great example of the efficiency of our model. More complex architectures that we tested do not necessarily provide better results. Also, the number of epochs is sufficient to get a good idea of each architecture's accuracy. Adam optimizer is used for training, and its learning rate is set to 0.04. The criterion is Cross Entropy Loss and the regularization factor is 2e-4. Dropout is applied to all feature vectors at a rate of 0.5, and dropout among GNN layers is set to 0.8.

After running the NAS method we process the results and examine the five architectures with the best accuracy in the validation set. Each architecture is now trained with 100 epochs on the TwiBot-20 dataset (Feng et al., 2021b). The train set is 70% of the dataset, the validation set is 20% and the test set is 10%. Adam optimizer with a learning rate of 1e-3 is also used for training. Then each architecture is tested on the test set. We will present the findings of these experiments below.

6.3 Evaluation metrics

We assess our model's performance using its Accuracy, F1-score, Precision, Recall, Specificity, and MCC. These metrics are computed by labeling the bots as the positive class and the genuine users as the negative class. To compare the performance of our model to the other baseline models we will only use the metrics Accuracy, F1-score, and MCC.

7 Results

Each architecture during the search is saved with its P-T configuration, accuracy in the validation set, and accuracy in the test set. In Figure 2, the five architectures with the highest validation accuracy that are chosen from the NAS method are depicted.

These architectures are trained and tested from scratch in TwiBot-20 dataset. We present all the metrics attained by all the architectures in Table 4.

All selections achieve good metrics and present advantages in bot detection over state-of-the-art methods. These results underscore the significant advantages that emerge from employing architecture search techniques regarding the field of bot recognition. Moreover, they establish the efficiency of utilizing user features and relationships between users in bot detection.

Upon closer examination of the results, the third architecture achieves the best evaluation metrics. The fifth architecture has the highest precision. However, all the architectures present high metrics of accuracy, F1-score, and MCC and whichever architecture we choose could compete with state-of-the-art models. From now on we will refer to the third architecture as our model, since it provides the highest accuracy.

In Table 5 we present the performance of the baseline methods on the TwiBot-20 dataset compared to ours. We see that our model benefits from the search for the fittest architecture that we performed beforehand, as it achieves a higher accuracy, F1-score, and MCC than other state-of-the-art methods.

8 Ablation study

To demonstrate our model's effectiveness and integrity we will perform an ablation study on the basic ideas: the user's features used for the training, the Gate operation, and the skip-connection operation.

To prove that using multi-modal information is vital to our model performance we will train the architecture that produces the best results with reduced features. We will reduce one feature at a time and use combinations of the features for the training. We present the results in Table 6.

We see that training with reduced features may achieve higher metrics in some cases. Notably, training without descriptions has a higher F1-score than the original model but has a lower precision. Also, training without tweets has a higher recall value. Training without numerical properties has a higher precision and specificity but a lower MCC than training without description. Training with only the categorical and numerical properties has the highest recall. Therefore, training with combinations of features does not achieve as high metrics as training with all the features in each case, meaning that all features contribute to the model's performance. These remarks are important to consider for future research in ensuring the dataset's quality, but training the model with all the features provided makes it more adaptable to other datasets. For further understanding we will train the model using only one feature at a time, to investigate their importance separately. We present the results in Table 7.

Obviously, the model trained with all the features has the best performance. From the results, we deduce that the categorical property is the feature that contributes the most to the model's sufficient accuracy. This ablation study proves that all features are advantageous for training our model to perform well in the task of bot detection. However, they do not contribute equally, and more

TABLE 4 Performance of the architectures from architecture search.

Model	Accuracy	F1-score	Precision	Recall	Specificity	MCC
1st Architecture	0.852 ± 0.005	0.865 ± 0.008	0.851 ± 0.015	0.880 ± 0.031	0.818 ± 0.027	0.702 ± 0.010
2nd Architecture	0.855 ± 0.004	0.869 ± 0.005	0.853 ± 0.007	0.886 ± 0.012	0.819 ± 0.012	0.709 ± 0.009
3rd Architecture	**0.857 ± 0.004**	**0.871 ± 0.003**	0.849 ± 0.008	**0.895 ± 0.007**	0.812 ± 0.013	**0.712 ± 0.007**
4th Architecture	0.852 ± 0.006	0.864 ± 0.008	0.856 ± 0.009	0.873 ± 0.026	0.828 ± 0.018	0.702 ± 0.013
5th Architecture	0.852 ± 0.007	0.864 ± 0.008	**0.858 ± 0.003**	0.872 ± 0.019	**0.829 ± 0.007**	0.703 ± 0.014

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

TABLE 5 Performance of models on the TwiBot-20 dataset.

Model	Accuracy	F1-score	MCC
Lee et al. (2021)	0.7456	0.7823	0.4879
Yang et al. (2020)	0.8191	0.8546	0.6643
Kudugunta and Ferrara (2018)	0.8174	0.7517	0.6710
Wei and Nguyen (2019)	0.7126	0.7533	0.4193
Miller et al. (2014)	0.4801	0.6266	-0.1372
Cresci et al. (2017)	0.4793	0.1072	0.0839
Davis et al. (2016)	0.5584	0.4892	0.1558
Ali Alhosseini et al. (2019)	0.6813	0.7318	0.3543
Feng et al. (2021a)	0.8412	0.8642	0.6863
Feng et al. (2021c)	0.8462	0.8707	0.7021
Ilias et al. (2024a)	0.7466	0.7630	–
Ours	**0.8568 ± 0.004**	**0.8712 ± 0.003**	**0.7116 ± 0.007**

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

studies to enhance the quality of the datasets could benefit future studies of bot detection.

Next, we compare the architecture that results from the architecture search with a Gate operation and without a Gate operation. The findings of this ablation study are depicted in Table 8. We see that the architecture without the gate has a reduced accuracy by 0.5% compared to the model's and a reduced F1-score by 0.46%. The gating mechanism dynamically consolidates information from all propagation steps, effectively regulating the smoothness of various nodes. Without it, the T operations take as input only the last output of the P steps. This is the reason the model underperforms without the Gate operation in the P functions, as it may suffer from over-smoothing. The architectures that are examined during this search have more T steps and shallower propagation processes, failing to obtain information from nodes during message passing as successfully as the original model. This ablation study proves the importance of the Gate operation in the P functions during our architecture search.

Finally, we compare the architecture that results from the architecture search with a skip-connection operation and without a skip-connection operation. The findings of this ablation study are depicted in Table 9. We see that the architecture without the gate has a reduced accuracy by 0.93% compared to the model's and a reduced F1-score by 1.2%. Without the skip-connection operation, the input of the T steps is only the output of the last step. This may lead to the degradation of the model as the transformation functions can increase. The processing of the messages from nodes is not as effective and the accuracy declines. This ablation study proves the importance of the skip-connection operation in the T functions during our architecture search.

9 Discussion

9.1 Implications

The proliferation of social media bots has prompted concerns regarding user safety and their broader societal impact. Bot detection, a focal point of contemporary studies, is not only explored through the lens of machine learning but also delves into the realms of social science. Various methodologies have been employed, encompassing supervised or unsupervised learning or a hybrid of both. A relatively recent and innovative approach involves Graph Neural Network (GNN)-based architectures, integrating diverse user features and interactions to construct a comprehensive graph representation. In our work, we formulate a heterogeneous graph that captures the following relationships between users, incorporating nodes with information on user profiles, tweets, and interests. This novel contribution enhances existing bot detection research by demonstrating the efficacy of integrating and analyzing user relationships.

As technology advances, the adaptive nature of bots poses an ongoing challenge for detection models, rendering many state-of-the-art architectures ineffective against newer datasets. The pressing need for adaptable models underscores the importance of overcoming the limitations associated with fixed architectures. Neural Architecture Search (NAS) models prove to be a promising solution, demonstrating their potential to enhance model efficiency in real-world tasks by automatically searching through various architectures. Historically, the adoption of NAS techniques for bot detection is limited, so we propose the implementation of an adapted DFG-NAS. By integrating DFG-NAS and tailoring it to Relational Graph Convolutional Networks, we explore optimal permutations of Propagation and Transformation steps in the message-passing protocol of the RGCN layers. Our investigation showcases superior performances of the top architectures compared to state-of-the-art models. Our work is one of the starting points in implementing architecture search

TABLE 6 Training model with less features.

Model	Accuracy	F1-score	Precision	Recall	Specificity	MCC
Ours	0.857 ± 0.004	0.871 ± 0.003	0.849 ± 0.008	0.895 ± 0.007	0.812 ± 0.013	0.712 ± 0.007
w/o description	**0.859 ± 0.004**	**0.875 ± 0.004**	0.845 ± 0.002	0.906 ± 0.008	0.804 ± 0.004	**0.718 ± 0.008**
w/o tweets	0.833 ± 0.007	0.858 ± 0.007	0.796 ± 0.004	0.93 ± 0.013	0.719 ± 0.005	0.671 ± 0.016
w/o numerical	**0.859 ± 0.003**	0.872 ± 0.005	**0.856 ± 0.012**	0.889 ± 0.023	**0.823 ± 0.021**	0.716 ± 0.007
w/o categorical	0.792 ± 0.003	0.814 ± 0.001	0.791 ± 0.010	0.840 ± 0.014	0.738 ± 0.021	0.582 ± 0.005
Des + tweets	0.759 ± 0.007	0.773 ± 0.009	0.789 ± 0.022	0.758 ± 0.034	0.761 ± 0.045	0.519 ± 0.014
Cat + num	0.817 ± 0.001	0.855 ± 0.001	0.749 ± 0.001	**0.996 ± 0.001**	0.607 ± 0.002	0.668 ± 0.002

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

TABLE 7 Training model with only one feature.

Model	Accuracy	F1-score	Precision	Recall	Specificity	MCC
Ours	**0.857 ± 0.004**	**0.871 ± 0.003**	**0.849 ± 0.008**	0.895 ± 0.007	**0.812 ± 0.013**	**0.712 ± 0.007**
only description	0.699 ± 0.007	0.74 ± 0.008	0.695 ± 0.015	0.793 ± 0.033	0.589 ± 0.046	0.392 ± 0.014
only tweets	0.585 ± 0.011	0.643 ± 0.017	0.602 ± 0.008	0.691 ± 0.037	0.461 ± 0.033	0.157 ± 0.022
only numerical	0.679 ± 0.02	0.758 ± 0.013	0.641 ± 0.018	0.929 ± 0.034	0.385 ± 0.063	0.383 ± 0.039
only categorical	0.817 ± 0.001	0.853 ± 0.001	0.747 ± 0.001	**1.000 ± 0.001**	0.6 ± 0.001	0.667 ± 0.001

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

TABLE 8 Ablation study on gate operation.

Model	Accuracy	F1-score	Precision	Recall	Specificity	MCC
With gate	**0.857 ± 0.004**	**0.871 ± 0.003**	**0.849 ± 0.008**	**0.895 ± 0.007**	**0.812 ± 0.013**	**0.712 ± 0.007**
Without gate	0.853 ± 0.003	0.867 ± 0.004	0.845 ± 0.010	0.891 ± 0.016	0.808 ± 0.018	0.704 ± 0.007

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

TABLE 9 Ablation study on skip-connection operation.

Model	Accuracy	F1-score	Precision	Recall	Specificity	MCC
With skip	**0.857 ± 0.004**	**0.871 ± 0.003**	0.849 ± 0.008	**0.895 ± 0.007**	0.812 ± 0.013	**0.712 ± 0.007**
Without skip	0.849 ± 0.009	0.860 ± 0.01	**0.857 ± 0.010**	0.863 ± 0.026	**0.831 ± 0.017**	0.695 ± 0.018

Values are reported as mean ± standard deviation. Five runs of results are averaged. The best outcomes for each evaluation metric are in bold.

models on bot detection. Our research findings encourage further exploration into how NAS models can automatically construct more effective architectures, resulting in a future restraint of the existence of bots.

9.2 Applicability of our approach to different types of social interaction

In this section, we examine the applicability of our introduced approach to other types of social interaction besides social media.

- Online gaming: bots impersonate human players to manipulate game outcomes. Bots are capable of playing without breaks. Therefore, they are able to gather resources, items, and so on very quickly which help them go to the next stage of gaming (Kang et al., 2013). Thus, people end playing with bots; so, it is impossible to win them. This fact entails serious issues, i.e., unfair gaming. Therefore, the early detection of bots in gaming is crucial, in order to ensure fair play in competitive and multiplayer games. Our method could be adapted by using response times, movement patterns, and time-series data as input features.
- Customer reviews and rating platforms: bots are often used for creating fake reviews and inflating rating in review platforms, including Amazon and Yelp. The main aim of bots is to promote specific products, restaurants, and so on. Our approach could be easily adapted to this case, since textual, timing, and user behavior features will be used.
- Digital voting and polling systems: bots are used to alter the results of Internet Polling (Mohammadi and Abbasimehr, 2010). Therefore, early recognition of bots in voting is crucial, so as to ensure reliable outcomes. Our method can be adapted by integrating features, such as IP addresses, voting patterns, and timing.

- Email and messaging systems: bots are responsible for spam and phishing. Early detection of bots is crucial for enhancing security. Features, including email headers, IP addresses, etc., must be incorporated in our study.

9.3 Limitations

Our study comes with some limitations. Firstly, we conducted our experiments only on one dataset, which does not ensure generalizabilty of our proposed approach. Therefore, in the future, we aim to test our method on TwiBot-22 dataset (Feng et al., 2024). Secondly, our method is based on the collection of labeled data. Obtaining labeled data is a difficult task. For this reason, unsupervised and self-supervised learning algorithms have been developed for addressing the issue of labels' scarcity. Applying unsupervised and self-supervised learning in conjunction with our approach is one of our future plans. Thirdly, we did not tune the hyperparameters due to limited access to GPU resources. Hyperparameter tuning ensures that optimal performance is obtained. Finally, we represented each user as a concatenation of features. Concatenation does not capture the inherent correlation of the different modalities. In the future, we aim to use multimodal fusion methods for constructing each user's representation (Ilias et al., 2022; Ilias and Askounis, 2023a; Chatzianastasis et al., 2023).

10 Conclusions and future work

As social media continues to play a pivotal role in shaping public opinion and discourse, the development of effective and adaptive bot detection methods becomes increasingly crucial for maintaining the integrity and trustworthiness of online information. In this study, we introduced a novel model for identifying bots, integrating GNNs and NAS algorithms, demonstrating significant performance gains. The integration of Graph Neural Architecture Search empowered us to dynamically determine optimal combinations of propagation and transformation operations in the graph neural network architecture. This adaptive architecture effectively addresses the constraints imposed by fixed structures, introducing a level of flexibility essential for improving the performance on the bot detection task. From the experiment results we conclude that the five architectures with the highest validation accuracy, during the architecture search, are quite efficient in our task and compete with other models. Meanwhile, the one with the highest accuracy achieves a test accuracy of 85.68%, surpassing other state-of-the-art models for bot detection. The outcomes of the experiment present promising prospects for integrating more Neural Architecture Search (NAS) methods into the domain of bot detection in various social media platforms.

The exploration of dynamic graph adaptations stands as a crucial avenue for future research in the task of bot identification in social media platform X. The dynamic nature of social networks, characterized by the continuous incorporation of new users, necessitates the development of mechanisms to seamlessly integrate these additions into the evolving graph structure. Investigating methods for real-time graph updates and exploring how the model adapts to the inclusion of new users will enhance the system's agility in capturing emerging bot behaviors within the dynamic social landscape. Furthermore, the prospect of transferring our model to other social media platforms emerges as a key future avenue. Extending the applicability of our approach beyond X involves understanding the unique dynamics and characteristics of different platforms. Future work should focus on developing a transferable framework capable of recognizing bot-like behaviors across diverse social networks. By addressing the nuances and variations in user interactions and content features, we can contribute to the development of a versatile bot detection system with broader applications in the ever-expanding realm of social media platforms.

Data availability statement

Publicly available datasets were analysed in this study. This data can be found here: https://github.com/BunsenFeng/TwiBot-20. Please contact shangbin@cs.washington.edu to obtain permission to download the dataset for research efforts only.

Author contributions

GT: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. MC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – review & editing. LI: Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing – review & editing. GK: Project administration, Supervision, Writing – review & editing. JP: Project administration, Resources, Supervision, Writing – review & editing. DA: Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, 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.

Publisher's note

References

Alarfaj, F. K., Ahmad, H., Khan, H. U., Alomair, A. M., Almusallam, N., Ahmed, M., et al. (2023). Twitter bot detection using diverse content features and applying machine learning algorithms. *Sustainability* 15:6662. doi: 10.3390/su15086662

Alauthman, M., Aslam, N., Al-kasassbeh, M., Khan, S., Al-Qerem, A., and Raymond Choo, K.-K. (2020). An efficient reinforcement learning-based botnet detection approach. *J. Netw. Comput. Appl.* 150:102479. doi: 10.1016/j.jnca.2019.102479

Ali Alhosseini, S., Bin Tareaf, R., Najafi, P., and Meinel, C. (2019). "Detect me if you can: spam bot detection using inductive representation learning," in *Companion Proceedings of The 2019 World Wide Web Conference, WWW '19* (New York, NY: Association for Computing Machinery), 148–153. doi: 10.1145/3308560.3316504

Alothali, E., Salih, M., Hayawi, K., and Alashwal, H. (2022). Bot-mgat: a transfer learning model based on a multi-view graph attention network to detect social bots. *Appl. Sci.* 12:8117. doi: 10.3390/app12168117

Alsmadi, I., and O'Brien, M. J. (2020). How many bots in Russian troll tweets? *Inf. Process. Manag.* 57:102303. doi: 10.1016/j.ipm.2020.102303

Bazmi, P., Asadpour, M., and Shakery, A. (2023). Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Inf. Process. Manag.* 60:103146. doi: 10.1016/j.ipm.2022.103146

Bessi, A., and Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday*, 21. doi: 10.5210/fm.v21i11.7090

Bui, T., and Potika, K. (2022). "Twitter bot detection using social network analysis," in *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, 87–88. doi: 10.1109/TransAI54797.2022.00022

Cai, S., Li, L., Deng, J., Zhang, B., Zha, Z.-J., Su, L., et al. (2021). "Rethinking graph neural architecture search from message-passing," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN: IEEE), 6653–6662. doi: 10.1109/CVPR46437.2021.00659

Chatzianastasis, M., Ilias, L., Askounis, D., and Vazirgiannis, M. (2023). "Neural architecture search with multimodal fusion methods for diagnosing dementia," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Rhodes Island: IEEE), 1–5. doi: 10.1109/ICASSP49357.2023.10096579

Chavoshi, N., Hamooni, H., and Mueen, A. (2016). "Debot: Twitter bot detection via warped correlation," in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (Barcelona: IEEE), 817–822. doi: 10.1109/ICDM.2016.0096

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Trans. Dependable Secure Comput.* 15, 561–576. doi: 10.1109/TDSC.2017.2681672

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). "Botornot: a system to evaluate social bots," in *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion* (Geneva: International World Wide Web Conferences Steering Committee), 273–274. doi: 10.1145/2872518.2889302

Dehghan, A., Siuta, K., Skorupka, A., Dubey, A., Betlen, A., Miller, D., et al. (2023). Detecting bots in social-networks using node and structural embeddings. *J. Big Data* 10:119. doi: 10.1186/s40537-023-00796-3

Dimitriadis, I., Dialektakis, G., and Vakali, A. (2024). Caleb: a conditional adversarial learning framework to enhance bot detection. *Data Knowl. Eng.* 149:102245. doi: 10.1016/j.datak.2023.102245

El-Mawass, N., Honeine, P., and Vercouter, L. (2020). Similcatch: enhanced social spammers detection on Twitter using markov random fields. *Inf. Process. Manag.* 57:102317. doi: 10.1016/j.ipm.2020.102317

Feng, S., Tan, Z., Li, R., and Luo, M. (2022). "Heterogeneity-aware Twitter bot detection with relational graph transformers," in *AAAI Conference on Artificial Intelligence, Vol. 36* (Vancouver, BC), 3977–3985. doi: 10.1609/aaai.v36i4.20314

Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., et al. (2024). "Twibot-22: towards graph-based Twitter bot detection," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22* (Red Hook, NY: Curran Associates Inc).

Feng, S., Wan, H., Wang, N., Li, J., and Luo, M. (2021a). "Satar: a self-supervised approach to Twitter account representation learning and its application in bot detection," in *Proceedings of the 30th ACM International Conference on Information; Knowledge Management, CIKM '21* (New York, NY: ACM), 3808–3817. doi: 10.1145/3459637.3481949

Feng, S., Wan, H., Wang, N., Li, J., and Luo, M. (2021b). "Twibot-20: a comprehensive Twitter bot detection benchmark," in *Proceedings of the 30th ACM International Conference on Information Knowledge Management, CIKM '21* (New York, NY: Association for Computing Machinery), 4485–4494. doi: 10.1145/3459637.3482019

Feng, S., Wan, H., Wang, N., and Luo, M. (2021c). "Botrgcn: Twitter bot detection with relational graph convolutional networks," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '21* (New York, NY: ACM), 236–239. doi: 10.1145/3487351.3488336

Ferrara, E. (2020). *What types of COVID-19 conspiracies are populated by Twitter bots?* First Monday. doi: 10.5210/fm.v25i6.10633

Gao, Y., Yang, H., Zhang, P., Zhou, C., and Hu, Y. (2021). "Graph neural architecture search," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20*, ed. C. Bessiere (International Joint Conferences on Artificial Intelligence Organization), 1403–1409. doi: 10.24963/ijcai.2020/195

Ilias, L., and Askounis, D. (2023a). Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowl.-Based Syst.* 277:110834. doi: 10.1016/j.knosys.2023.110834

Ilias, L., and Askounis, D. (2023b). Multitask learning for recognizing stress and depression in social media. *Online Soci. Netw. Media* 37–38: 100270. doi: 10.1016/j.osnem.2023.100270

Ilias, L., Askounis, D., and Psarras, J. (2022). "A multimodal approach for dementia detection from spontaneous speech with tensor fusion layer," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (Ioannina: IEEE), 1–5. doi: 10.1109/BHI56158.2022.9926818

Ilias, L., Michail Kazelidis, I., and Askounis, D. (2024a). Multimodal detection of bots on x (Twitter) using transformers. *IEEE Trans. Inf. Forensics Secur.* 19, 7320–7334. doi: 10.1109/TIFS.2024.3435138

Ilias, L., Mouzakitis, S., and Askounis, D. (2024b). Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Trans. Comput. Soc. Syst.* 11, 1979–1990. doi: 10.1109/TCSS.2023.3283009

Ilias, L., and Roussaki, I. (2021). Detecting malicious activity in Twitter using deep learning techniques. *Appl. Soft Comput.* 107:107360. doi: 10.1016/j.asoc.2021.107360

Jiang, S., and Balaprakash, P. (2020). "Graph neural network architecture search for molecular property prediction," in *2020 IEEE International Conference on Big Data (Big Data)* (Los Alamitos, CA: IEEE Computer Society), 1346–1353. doi: 10.1109/BigData50022.2020.9378060

Kang, A. R., Woo, J., Park, J., and Kim, H. K. (2013). Online game bot detection based on party-play log analysis. *Comp. Math. Appl.* 65, 1384–1395. doi: 10.1016/j.camwa.2012.01.034

Kerasiotis, M., Ilias, L., and Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Soc. Netw. Anal. Mining* 14:196. doi: 10.1007/s13278-024-01360-4

Koggalahewa, D., Xu, Y., and Foo, E. (2022). An unsupervised method for social network spammer detection based on user information interests. *J. Big Data* 9:7. doi: 10.1186/s40537-021-00552-5

Kudugunta, S., and Ferrara, E. (2018). Deep neural networks for bot detection. *Inf. Sci.* 467, 312–322. doi: 10.1016/j.ins.2018.08.019

Kušen, E., and Strembeck, M. (2020). You talkin' to me? Exploring human/bot communication patterns during riot events. *Inf. Process. Manag* 57:102126. doi: 10.1016/j.ipm.2019.102126

Lee, K., Eoff, B., and Caverlee, J. (2021). Seven months with the devils: a long-term study of content polluters on Twitter. *Proc. Int. AAAI Conf. Web Soc. Media* 5, 185–192. doi: 10.1609/icwsm.v5i1.14106

Li, Y., Wen, Z., Wang, Y., and Xu, C. (2021). One-shot graph neural architecture search with dynamic search space. *Proc. AAAI Conf. Artif. Intell.* 35, 8510–8517. doi: 10.1609/aaai.v35i10.17033

Li, Y., Wu, J., and Deng, T. (2023). Meta-gnas: meta-reinforcement learning for graph neural architecture search. *Eng. Appl. Artif. Intell.* 123:106300. doi: 10.1016/j.engappai.2023.106300

Liu, Y., Tan, Z., Wang, H., Feng, S., Zheng, Q., Luo, M., et al. (2023). "Botmoe: Twitter bot detection with community-aware mixtures of modal-specific experts," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23* (New York, NY: Association for Computing Machinery), 485–495. doi: 10.1145/3539618.3591646

Lopes, D. A. G., Marotta, M. A., Ladeira, M., and Gondim, J. J. C. (2022). "Botnet detection based on network flow analysis using inverse statistics," in *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)* (Madrid: IEEE), 1–6. doi: 10.23919/CISTI54924.2022.9820318

Lopes, V., Santos, M., Degardin, B., and Alexandre, L. A. (2024). Guided evolutionary neural architecture search with efficient performance estimation. *Neurocomputing* 584:127509. doi: 10.1016/j.neucom.2024.127509

Mahmud, T., Ptaszynski, M., Eronen, J., and Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: review of the state of the art. *Inf. Process. Manag.* 60:103454. doi: 10.1016/j.ipm.2023.103454

Mannocci, L., Cresci, S., Monreale, A., Vakali, A., and Tesconi, M. (2022). "Mulbot: unsupervised bot detection based on multivariate time series," in *2022 IEEE*

International Conference on Big Data (Big Data) (Los Alamitos, CA: IEEE Computer Society), 1485–1494. doi: 10.1109/BigData55660.2022.10020363

Mi, Y., and Apuke, O. D. (2024). How does social media knowledge help in combating fake news? Testing a structural equation model. *Think. Skills Creat.* 52:101492. doi: 10.1016/j.tsc.2024.101492

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., and Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Inf. Sci.* 260, 64–73. doi: 10.1016/j.ins.2013.11.016

Minnich, A., Chavoshi, N., Koutra, D., and Mueen, A. (2017). "Botwalk: efficient adaptive exploration of Twitter bot networks," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (New York, NY: ACM), 467–474. doi: 10.1145/3110025.3110163

Mohammadi, S., and Abbasimehr, H. (2010). "A high level security mechanism for internet polls," in *2010 2nd International Conference on Signal Processing Systems, Volume 3* (Dalian: IEEE), V3-101–V3-105. doi: 10.1109/ICSPS.2010.5555837

Noekhah, S. binti Salim, N., Zakaria, N. H. (2020). Opinion spam detection: using multi-iterative graph-based model. *Inf. Process. Manag.* 57:102140. doi: 10.1016/j.ipm.2019.102140

Nunes, M., and Pappa, G. L. (2020). *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*. Cham: Springer International Publishing.

Peng, W., Hong, X., Chen, H., and Zhao, G. (2020). Learning graph convolutional network for skeleton-based human action recognition by neural searching. *Proc. AAAI Conf. Artif. Intell.* 34, 2669–2676. doi: 10.1609/aaai.v34i03.5652

Peng, Y., Song, A., Ciesielski, V., Fayek, H., and Chang, X. (2023). "Fast evolutionary neural architecture search by contrastive predictor with linear regions," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '23* (New York, NY: Association for Computing Machinery), 1257–1266. doi: 10.1145/3583131.3590452

Pham, P., Nguyen, L. T., Vo, B., and Yun, U. (2022). Bot2vec: a general approach of intra-community oriented representation learning for bot detection in different types of social networks. *Inf. Syst.* 103:101771. doi: 10.1016/j.is.2021.101771

Quezada, V., Astudillo-Salinas, F., Tello-Oquendo, L., and Bernal, P. (2023). Real-time bot infection detection system using dns fingerprinting and machine-learning. *Comput. Netw.* 228:109725. doi: 10.1016/j.comnet.2023.109725

Saxena, N., Sinha, A., Bansal, T., and Wadhwa, A. (2023). A statistical approach for reducing misinformation propagation on Twitter social media. *Inf. Process. Manag.* 60:103360. doi: 10.1016/j.ipm.2023.103360

Scheibenzuber, C., Neagu, L.-M., Ruseti, S., Artmann, B., Bartsch, C., Kubik, M., et al. (2023). Dialog in the echo chamber: fake news framing predicts emotion, argumentation and dialogic social knowledge building in subsequent online discussions. *Comput. Human Behav.* 140:107587. doi: 10.1016/j.chb.2022.107587

Shang, R., Zhu, S., Liu, H., Ma, T., Zhang, W., Feng, J., et al. (2023). Evolutionary architecture search via adaptive parameter control and gene potential contribution. *Swarm Evol. Comput.* 82:101354. doi: 10.1016/j.swevo.2023.101354

Shevtsov, A., Antonakaki, D., Lamprou, I., Pratikakis, P., and Ioannidis, S. (2023). Botartist: Twitter bot detection machine learning model based on Twitter suspension. *arXiv* [Preprint]. arXiv:2306.00037. doi: 10.48550/arXiv.2306.00037

Sujith, K., Chowdhury, S., Goyal, A., Hegde, A. V., and Srinath, R. (2022). "Twitter bot detection and ranking using supervised machine learning models," in *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Volume 01* (Chennai: IEE), 1–6. doi: 10.1109/ICDSAAI55433.2022.10028860

Uyheng, J., Moffitt, J., and Carley, K. M. (2022). The language and targets of online trolling: a psycholinguistic approach for social cybersecurity. *Inf. Proc. Manag.* 59:103012. doi: 10.1016/j.ipm.2022.103012

Wei, F., and Nguyen, U. T. (2019). "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (Los Angeles, CA: IEEE), 101–109. doi: 10.1109/TPS-ISA48467.2019.00021

Wei, F., and Nguyen, U. T. (2023). Twitter bot detection using neural networks and linguistic embeddings. *IEEE Open J. Comput. Soc.* 4, 218–230. doi: 10.1109/OJCS.2023.3302286

Wu, J., Teng, E., and Cao, Z. (2022). "Twitter bot detection through unsupervised machine learning," in *2022 IEEE International Conference on Big Data (Big Data)* (Osaka: IEEE), 5833–5839. doi: 10.1109/BigData55660.2022.10020983

Xu, Y., Zhou, D., and Wang, W. (2023). Being my own gatekeeper, how i tell the fake and the real – fake news perception between typologies and sources. *Inf. Process. Manag.* 60:103228. doi: 10.1016/j.ipm.2022.103228

Yang, K.-C., Ferrara, E., and Menczer, F. (2022). Botometer 101: social bot practicum for computational social scientists. *J. Comput. Soc. Sci.* 5, 1511–1528. doi: 10.1007/s42001-022-00177-5

Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. *Proc. AAAI Conf. Artif. Intell.* 34, 1096–1103. doi: 10.1609/aaai.v34i01.5460

Yang, Y., Yang, R., Li, Y., Cui, K., Yang, Z., Wang, Y., et al. (2023). Rosgas: adaptive social bot detection with reinforced self-supervised gnn architecture search. *ACM Trans. Web* 17, 1–31. doi: 10.1145/3572403

Ye, S., Tan, Z., Lei, Z., He, R., Wang, H., Zheng, Q., et al. (2023). Hofa: Twitter bot detection with homophily-oriented augmentation and frequency adaptive attention. *arXiv* [Preprint]. arXiv:2306.12870. doi: 10.48550/arXiv.2306.12870

Zhang, W., Lin, Z., Shen, Y., Li, Y., Yang, Z., Cui, B., et al. (2022). "Deep and flexible graph neural architecture search," in *Proceedings of the 39th International Conference on Machine Learning, Volume 162 of Proceedings of Machine Learning Research*, eds. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (Baltimore, MA: PMLR), 26362–26374. Available at: https://proceedings.mlr.press/v162/zhang22s.html

Zhang, Y., Ma, J., and Fang, F. (2024). How social bots can influence public opinion more effectively: Right connection strategy. *Phys. A Stat. Mech. Appl.* 633:129386. doi: 10.1016/j.physa.2023.129386

Zhao, H., Wei, L., and Yao, Q. (2020). Simplifying architecture search for graph neural network. *arXiv* [Preprint]. arXiv:2008.11652. doi: 10.48550/arXiv.2008.11652

Zhao, H., Yao, Q., and Tu, W. (2021). "Search to aggregate neighborhood for graph neural network" in *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (Chania: IEEE), 552–563. doi: 10.1109/ICDE51399.2021.00054

Zhao, J., Zhang, R., Zhou, Z., Chen, S., Jin, J., Liu, Q., et al. (2021). A neural architecture search method based on gradient descent for remaining useful life estimation. *Neurocomputing* 438, 184–194. doi: 10.1016/j.neucom.2021.01.072

Zhou, K., Huang, X., Song, Q., Chen, R., and Hu, X. (2022). Auto-gnn: neural architecture search of graph neural networks. *Front. Big Data* 5:1029307. doi: 10.3389/fdata.2022.1029307

frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 07 January 2025
DOI 10.3389/frai.2024.1474034

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Yenny Villuendas Rey,
Instituto Politécnico Nacional, Mexico
Jonathan Hardy,
University of the Arts London,
United Kingdom
Bernard Siman,
Egmont- Royal Institute for International Relations, Belgium

*CORRESPONDENCE
Anatolii Marushchak
✉ amarushchak@ukr.net

RECEIVED 31 July 2024
ACCEPTED 18 December 2024
PUBLISHED 07 January 2025

CITATION
Marushchak A, Petrov S and Khoperiya A (2025) Countering AI-powered disinformation through national regulation: learning from the case of Ukraine.
Front. Artif. Intell. 7:1474034.
doi: 10.3389/frai.2024.1474034

Countering AI-powered disinformation through national regulation: learning from the case of Ukraine

Anatolii Marushchak[1]*, Stanislav Petrov[2] and Anayit Khoperiya[3]

[1]International Information Academy, Kyiv, Ukraine, [2]The Institute of Special Communication and Information Protection of the National Technical University of Ukraine "Ihor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, [3]Deputy Head of the Center for Countering Disinformation of the National Security and Defense Council of Ukraine, Kyiv, Ukraine

Advances in the use of AI have led to the emergence of a greater variety of forms disinformation can take and channels for its proliferation. In this context, the future of legal mechanisms to address AI-powered disinformation remains to be determined. Additional complexity for legislators working in the field arises from the need to harmonize national legal frameworks of democratic states with the need for regulation of potentially dangerous digital content. In this paper, we review and analyze some of the recent discussions concerning the use of legal regulation in addressing AI-powered disinformation and present the national case of Ukraine as an example of developments in the field. We develop the discussion through an analysis of the existing counter-disinformation ecosystems, the EU and US legislation, and the emerging regulations of AI systems. We show how the Ukrainian Law on Counter Disinformation, developed as an emergency response to internationally recognized Russian military aggression and hybrid warfare tactics, underscores the crucial need to align even emergency measures with international law and principles of free speech. Exemplifying the Ukrainian case, we argue that the effective actions necessary for countering AI-powered disinformation are prevention, detection, and implementation of a set of response actions. The latter are identified and listed in this review. The paper argues that there is still a need for scaling legal mechanisms that might enhance top-level challenges in countering AI-powered disinformation.

KEYWORDS

disinformation, artificial intelligence, law, regulation, prevention, detection, response

1 Introduction

Over the past decade, digital technology and media have advanced enormously. While these developments benefit individuals and societies, driving economic and social development, they also open the door to information manipulation at scale. Benefitting from the pervasive use of social media (SM) worldwide, perpetrators systematically spread disinformation to destabilize societies, interfere in state governance, and radicalize groups. Addressing these advancements and their impact on states and individuals is a pressing issue for legislators at both national and international levels.

The complexity inherent to countering disinformation campaigns from the legal perspective stems from the juxtaposition between the need to regulate cyber operations and the limits of applying regulations and restrictions on freedom of speech. It is not yet clear what regulations are to be applied to cyber operations (including disinformation), especially under the existing international law frameworks.

On the one hand, such operations could be viewed as a part of espionage activities, which are not directly prohibited by international law (Rosen et al., 2023). On the other hand, they could be assessed as a breach of national sovereignty and be sanctioned under national and international law. There is another element adding to the complexity of the issue – given that democratic states constitutionally guarantee access to information and freedom of speech to their citizens, the extent of law applicability in the outlined scenario is not clear. Unrestricted access to information lies at the core of democratic regimes, as it is believed to enable their citizens to participate freely and fairly in the politics and civic life of the nation. At the same time, some of the existing international law frameworks provide the possibility of developing and applying legislative measures to counteract disinformation. For instance, the General Comment on Article 19 of Civil and Political Rights (ICCPR) provides that "When a State party invokes a legitimate ground for restriction of freedom of expression, it must demonstrate in specific and individualized fashion the precise nature of the threat and the necessity and proportionality of the specific action taken" (Refworld, 2024a).

Given that disinformation campaigns are disseminated predominantly through SM, special attention must be paid to this domain. The threat of disinformation for national security is growing because of perpetrators' use of artificial intelligence (AI) tools to create and disseminate false and manipulative messages. In particular, fake news websites, AI-generated personalities, and fraudulent accounts are all used to spread harmful narratives. AI-powered social bots can sense, think, and act on SM platforms similar to humans (Hajli et al., 2022).

At the moment, SM regulation is in incipient stages and varies from state to state. It mostly relies on the legal measures developed in the countries of origin (US or China) over the activity of the very large online platforms (VLOPs). According to Stiglitz, a Nobel laureate in economics and a professor at Columbia University "Big Tech's trade-pact ploy is to create a global digital architecture where America's digital giants can continue to dominate abroad and are unfettered at home and elsewhere (Stiglitz, 2024). The complexity of VLOPs' corporate regulation is an additional constraint for national governments to address disinformation on SM. Tackling AI-powered disinformation campaigns requires a multipronged approach involving cyber operations and freedom of speech regulations as well as AI-specific legislation and regulation.

This illustrates further that the constraints to addressing the aftermath of disinformation are that in democratic societies, legislation lags far behind the innovation of emerging technology due to the need for consensus decision-making and the lack of technical expertise possessed by legislators. The mitigation of disinformation and its consequences for national security is dependent on freedom of speech guarantees as well as privacy protection regulations. Still, states must ensure that media and SM are free from malign interference and that civil society participates in public space without disinformation, by enacting mechanisms that distinguish the truth from fiction.

Researchers contributed to defining the most appropriate division of the responsibilities between governments, industry, and civil society while addressing disinformation. Namely, Hamilton (2021) recognizes legal exemptions from fundamental freedom of speech based upon National Security concerns, analyzing the existing practice of content moderation. She defined "the modern free speech triangle" (nation-states, SM companies and users) in the context of responsibility for online content production, amplification, and rule creation and enforcement. Others (Peukert, 2024) pay attention to the justification and the challenges posed by anti-disinformation measures, including the current regulation of counter-disinformation in the EU and the US. Comprehensive analysis of the emerging EU anti-disinformation framework based on the Code of Practice on Disinformation and the Digital Services Act, aimed at minimizing the distribution of false or misleading information has also been conducted (Cavaliere, 2022).

Additionally, scientists identify two strategies for bolstering global AI governance in light of these collaboration issues, which are particularly insightful for our research. These are (a) creating new, centralized international AI institution(s) and (b) enhancing the capacities and coordination of already-existing organizations (Roberts et al., 2024). It has been argued (Roberts et al., 2024), that it is more politically acceptable and practical to fortify the weak "regime complex" of international organizations as they currently stand. Some concerns related to these technologies can be mitigated by inclusive and mutually reinforcing policy change, which in turn would be supported by improved coordination and capabilities amongst the current international organizations controlling AI.

In practical terms, the urgency of the issues introduced above is exemplified at least since Russia's annexation of Crimea. In the context of Russia's full-scale invasion of Ukraine, the Ukrainian government and international experts (Tregubov, 2021) recognize that Russian agents spread the most destructive disinformation developed with AI tools, also known as "real-time" deepfakes. The Atlantic Council, a think tank based in Washington DC, underscores that "the quest to find the right balance between free speech and security will shape and define the decades ahead… and Ukraine's experience should certainly be part of this global conversation." In this review, our aim is to provide the Ukrainian approach to legal regulations to counter Russian disinformation campaigns, particularly amplified by the use of AI.

This way, the goal of the paper is to address two concerns. The first pertains to the discrepancy between the pace at which AI technology is advancing and the pace at which national government apparatuses can respond, often hampered by the legal provisions inherent in national laws. The second is to discuss the nexus of security concerns and other legal principles, especially in democracies like Ukraine to address AI-powered disinformation.

The paper is divided into three main parts. Section 1 is dedicated to the interconnections between AI and disinformation. Here, both AI usage to spread disinformation and AI-based solutions to address disinformation are discussed. In Section 2 the paper identifies the challenges arising from counter-disinformation and AI regulations in EU, US, and Ukraine with additional emphasis on Foreign Information Manipulation and Interference (FIMI). Section 3 maps the law's applicability in AI and counter disinformation nexus. The self-regulation or binding legal regulations approaches for VLOPs are examined. Section 3 also proposes a set of preventative, detection, and response actions to address AI-powered disinformation.

2 AI and disinformation: interconnections

2.1 How AI is used to spread disinformation

As modern reality shows, AI can be both a tool for objective information reaching the masses and a powerful tool for spreading

false or manipulative messages. This opens up wide opportunities for those who intend to manipulate public opinion, control the information space, and influence political processes. People can become victims of disinformation without the ability to distinguish truth from manipulation. This is especially dangerous for political and public discourse, where the accuracy of the information can determine the future of a country and its citizens. For the Ukrainian government tackling Russian disinformation sometimes even means saving the lives of the country's citizens.

AI tools are actively used to exert destructive information influence. In particular, fake news websites or voiced AI-generated virtual personalities are used to spread manipulative information. Fake accounts of non-existent people are also being created to promote information needed by the perpetrator. AI can provide more comprehensible and accurate information than people, but it can also generate more persuasive disinformation (Spitale et al., 2023).

In recent years, deepfakes have often been used to exert destructive influence. Ukrainian government notes that Russian propaganda spreads the most dangerous type of deepfakes, called "real-time" deepfakes. "Real-time" deepfake technology poses a serious threat to the information sphere due to its ability to create fake videos instantaneously, potentially deceiving both the public and political elites of various countries, thereby influencing decision-making processes. This technology allows for quick and almost undetectable changes to content, including the faces of politicians or other influential figures, manipulating words and images. In light of such deep-fake capabilities, the threat of trust in information becomes critical for society.

One example of effective use of "real-time" deepfake technology is the propaganda show "Show ViL" by Russian pranksters "Vovan" (Vladimir Kuznetsov) and "Lexus" (Aleksei Stolyarov), known for their conversations with high-ranking officials from various countries. One of their notable uses of "real-time" deepfake was in a conversation with Krišjānis Kariņš, the former Prime Minister of Latvia (Rutube, 2024a) or former President of Poland Aleksander Kwaśniewski (Rutube, 2024b), where they discussed controversial and sensitive political and geopolitical issues. They used "real-time" deepfake technology to make video calls, pretending to be political figures from certain African countries.

The pranksters conduct video calls with the targeted persons on behalf of other public figures whose images are generated online by AI. Moreover, the generated images and sound are of such high quality that they do not raise any doubts among the victims of the prank. The content obtained in this way is published by the Russian side in the public domain to discredit the persons who became the target of such a propaganda "prank."

Ukrainian Center for Countering Disinformation (CCD), the working body of the National Security and Defense Council of Ukraine (NSDC) uncovered a damaging information campaign against President Volodymyr Zelenskyy constructed with the help of AI. The report "Information Influence Campaign in the African Information Space" (Center for Countering Disinformation, 2024) provides details of the campaign and is a striking illustration of foreign information manipulation and interference (FIMI). Using resources from African media, the campaign's primary objective was to denigrate Zelenskyy and Ukraine in the eyes of international allies. The quasi-state actors disseminated several bogus films and articles under false names through certain African media sources with anti-Ukrainian rhetoric.

A prominent case of this campaign is about the alleged ownership of a villa in Egypt by President Zelenskyy's family. On August 22, 2023, the Nigerian media outlet Punch published an article titled "A Luxury Villa Owned by President Zelenskyy's Family Found on the Coast of Egypt." The material, citing an investigation by "Egyptian journalist" Mohammed Al-Alawi from August 20, 2023, mentioned a villa in the Egyptian resort town of El Gouna, allegedly belonging to Zelenskyy's mother-in-law, Olga Kiyashko. However, the CCD found that the Zelenskyy family does not own any property in Egypt. Moreover, no evidence was found to confirm the existence of Mohammed Al-Alawi, indicating a high probability of AI tools being used to create this persona. Additionally, Mohammed Al-Alawi's YouTube channel (Center for Countering Disinformation, 2024) creation and the video about the villa were dated the same day. This also indicates manipulative tactics.

The dangers of AI technologies in spreading disinformation should be considered broadly, factoring in not only the creation and manipulation of content but also the use of AI to disseminate it, amplify the likelihood of preexisting threats, and profile users with greater precision. AI systems are currently being abused in several fields, and if they are employed more widely, there will be greater opportunities for abuse. On such misuses, decision-makers will feel obliged to step in, but it can be challenging to select the best set of responses (Anderljung and Hazell, 2023). Developers would have a strong incentive to set up organizational procedures for guaranteeing honest and efficient reporting if regulations imposed legal penalties for careless or intentional misreporting. Regulators-approved independent auditors may also be able to help find instances of misreporting (Kolt et al., 2024).

However, it might still be challenging to match the fundamental rights criteria with the decision models of more sophisticated algorithms (Buiten, 2019). Policymakers ought to concentrate on the dangers that they wish to lower. It is demonstrated that defining the primary sources of relevant risks – specific technological strategies (like reinforcement learning), applications (like facial recognition), and capabilities (like the capacity to engage physically with the environment) – better satisfies the requirements for legal definitions (Schuett, 2023).

The legal system faces both conceptual and practical issues as a result of the distinctive qualities of AI and how it can be developed (Scherer, 2015). According to Viljanen and Parviainen (2022), the five layers of AI law are the following: data rules that govern data use, application-specific rules that target AI applications or application domains, general AI rules that apply to a broad range of AI applications, application-specific non-AI rules that apply to specific activities but not to AI specifically, and general non-AI rules that apply generally and across domains. The last two layers are counter-disinformation legislation in our case.

Worries about AI safety have arisen because of AI systems' unpredictability, explainability, and uncontrollability. Because of the complexity of AI systems, limitations in human understanding, and elusiveness of emergent behaviors, it is impossible to predict certain capabilities with any degree of accuracy (Yampolskiy, 2024). Moreover, gaining an awareness of various rule complexes, their dynamics, and regulatory modalities is necessary to comprehend the regulatory environment around AI (Viljanen and Parviainen, 2022).

Governments must recognize the significance of adopting a regulatory framework that optimizes AI's advantages while accounting for its hazards. This might entail classifying and categorizing risks according to relevant legal frameworks and country situations, when applicable (AI Safety Summit, 2023).

Some scientists find compute governance to be a significant approach to AI governance. Tens of thousands of sophisticated AI chips are needed to train sophisticated AI systems; these chips cannot be purchased or used covertly. AI chips can be supplied to or taken away from specific actors and in certain situations due to their physical nature. Moreover, it is measurable: it is possible to measure chips, their attributes, and their utilization. The extremely concentrated structure of the AI supply chain further enhances the compute's detectability and excludability (Heim et al., 2024). Finally, if mitigations are not implemented for AI-based disinformation including legal regulations, interactive and compositional deepfakes have the potential to bring us closer to a post-epistemic world in which it will be impossible to tell fact from fiction (Horvitz, 2022).

2.2 AI-based solutions to address disinformation

The issue of countering AI-powered disinformation is extremely important, but it is greatly complicated because of the rapid development of the technology for generating deepfakes, including mentioned "real-time" deepfakes. Today, we know for certain about digital visual evidence that may indicate interference with AI content: occlusions, imperfect edges of visual masks, color and light mismatch, etc. However, it is predicted that these image defects will be eliminated in the near future, as AI technologies are developing extremely dynamically. A wide range of significant and urgent threats associated with AI are being discussed more and more by AI specialists, journalists, policymakers, and the public (Center for AI Safety, 2024).

There are already technical, legal, regulatory, and educational approaches to counter disinformation in Ukraine – some of which have already been implemented and some of which are just emerging – that can help reduce the level of threat associated with the use of AI. Looking at the case of Ukraine, it is worth paying attention to the activities of some of Ukrainian AI-based platforms (Osavul, 2024) which effectively help to detect destructive information influence campaigns in their early stages. Their capabilities are powered by CommSecure and CIB Guard software. CommSecure enables to detect specific narratives in messages on social networks and communities, such as public groups in messengers. This ensures that potentially dangerous information flows are quickly identified and analyzed. CIB Guard, on the other hand, specializes in analyzing public user pages, identifying bots, and determining whether they act in a coordinated manner. This approach allows to quickly recognize coordinated campaigns that may be aimed at manipulating public opinion or spreading disinformation.

3 Counter disinformation and AI regulations

First of all, we would like to mention that the current international law principle of sovereignty clashes with the cross-border nature of cyberspace and disinformation operations. The UN Open-Ended Working Groups (OEWG) as the multi-lateral forum for cyber diplomacy has not yet provided the recommendations applicable for countering the AI-powered disinformation.

Historically, one of the first cases of legally defined misconduct in disinformation is dated January 2019. Then the US a company that created fake SM profiles to make millions of dollars in revenue settled a case with the New York state attorney. The settlement is the first case in which law enforcement has concluded that selling fake SM activity is illegal (Funke and Flamini, 2024).

As stated above, VLOPs and their platforms are usually to some extent dependent on the laws of the host countries. However, their terms of service are devised based on the legal system of the country of origin, currently predominantly the US and China. VLOPs' intention to develop internal counter-disinformation mechanisms including AI-powered solutions can be affected by commercial and geopolitical apprehensions, e.g., risks of retributory regulation or losing access to market in the host country.

Contrary to the approach to making VLOPs responsible for addressing the disinformation Hamilton's (2021) opinion is that "as a strictly legal matter, there is no reason for the platforms to have developed the elaborate content moderation systems they currently run." VLOPs faced the risks of "wasting" time, finance, and human resources on addressing disinformation by monitoring their networks, detecting fake news and even losing their users if the moderation gives rise to public debate. Thus, VLOPs used to be reluctant to identify perpetrators of disinformation. Nevertheless, VLOPs under pressure or in collaboration with governments, predominantly the US and China, started to develop detection and suspension initiatives, including those relying on AI, aimed at bots and botnets, as well as users exposed to disinformation, reinforcing the visibility of reliable content produced by trustworthy media and fact-checking sources, and vice versa reducing visibility (Santa Clara University, 2024) or suspension of sites' disinformation content.

At this point various legal approaches to counter disinformation have been put in place, helping reduce the level of threat associated with the use of AI. According to many reports, legislation governing AI is still in its infancy, with few statutes and other regulatory tools governing the creation and application of AI (Viljanen and Parviainen, 2022). The traditional conundrum of defining AI is exacerbated by the fact that our knowledge of natural intelligence is still incomplete (Mahler, 2021). Overcoming the present shortcomings in global AI governance is complicated by first-order cooperation issues resulting from interstate competition and second-order cooperation issues arising from dysfunctional international institutions (Roberts et al., 2024).

Recently several declarations have been produced at the international level. For instance, UK Bletchley Park hosted the first global summit on frontier AI safety (Artificial intelligence, 2023). NATO Washington Summit Declaration adopted on 10 July 2024 also mentioned the intention of the NATO member-states to develop individual and collective capacity to analyze and counter hostile disinformation operations (NATO, 2024). Additionally, in May 2024 at the AI Safety Summit in Seoul the following AI businesses pledged to uphold a set of international guidelines for AI safety known as the Frontier AI Safety Commitments: Amazon, Anthropic, Cohere, Google, IBM, Inflection AI, Meta, Microsoft, Mistral AI, Open AI, Samsung, – Technology Innovation Institute, xAi, Zhipu.ai (Zhipu.ai

is a Chinese company backed by Alibaba, Ant and Tencent) (AI Seoul summit, 2024).

Finally, authorities like Ukrainian CCD to counteract disinformation have also been developed along regional lines. They've also established the basis of coalitions with such entities as the EU Intelligence and Situation Centre (INTCEN), EU East StratCom Task Force with the flagship project EUvsDisinfo, the Helsinki European Centre of Excellence for Countering Hybrid Threats (counter disinformation capacity), and the NATO Strategic Communication Excellence Centre, among other.

3.1 The EU regulations

Holding EU candidate status Ukraine is actively harmonizing its legislation with EU legislation and regulations, particularly in measures to counter AI-powered disinformation. The need to legally address the threat is based upon recent EU legal decisions. For instance, advanced disinformation/influence operations campaigns coupled with abuse of AI are listed as a revised line-up of the emerging cybersecurity threats to have an impact by 2030 in the EU (ENISA, 2024). Additionally, the European Union Council on May 21, 2024, approved two documents pertaining to disinformation management – the Future of EU Digital Policy (Council of the European Union, 2024a) and Council conclusions on democratic resilience: safeguarding electoral processes from foreign interference (Council of the European Union, 2024b). The first document seeks to establish the framework for the next 5 years of digital policymaking and disinformation is listed as one of the detrimental or illegal occurrences that must be combated while promoting entrepreneurship, innovation, and the growth of the capital market. The durability of democracy and preventing outside intervention in electoral processes are the main topics of the second document (Council of the European Union, 2024b). It also provides a comprehensive overview of the legislative, non-legislative, and institutional tools that the EU has established. Both these documents stress the importance of further legal developments in counter disinformation realm. Consequently, they will be reflected to some extent in Ukrainian regulation.

3.1.1 European media freedom act

Some norms of the European Media Freedom Act (EMFA) (European Media Freedom Act, 2024) will be harmonized with the Ukrainian legislation as well. As Gamito (2023) stated before enacting EMFA the "online media freedom" was regulated domestically due to the threats of disinformation, without having in mind the European internal market. From the Ukrainian perspective, particularly interesting will be powering the European Board for Media Services (the Board) to engage in dialogue with VLOP to monitor adherence to self-regulatory initiatives aiming to protect users from harmful content, including FIMI. The EMFA emphasizes "insufficient tools for regulatory cooperation between national regulatory authorities or bodies."

With reference to the EU Code of Practice on Disinformation (European Commission, 2024) the EMFA obliges the Board to organize a structured dialogue between providers of VLOPs, representatives of media service providers, and representatives of civil society. This is to foster access to diverse offerings of independent media including as regards the moderation processes by VLOPs and monitor adherence to self-regulatory initiatives to protect users from harmful content, including disinformation and FIMI (Gamito, 2023). The implementation of these norms into Ukrainian legislation is of vital importance to counter Russian AI-powered disinformation at VLOPs platforms.

3.1.2 European AI act

Another major document is European AI Act (European Parliament, 2024) that contains several provisions on disinformation that need to be reflected in Ukrainian legislation. The European AI Act[1] directly addresses systemic risks posed by general-purpose AI models. These risks include the risks from the facilitation of disinformation. Legally significant is the AI Act definition of "deepfake"[2] as AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful (European Parliament, 2024).

Moreover, the European AI Act[3] stipulates the obligations placed on providers and deployers of certain AI systems to enable the detection and disclosure that the outputs of those systems are artificially generated in particular as regards the obligations of providers of VLOPs or very large online search engines to identify and mitigate the dissemination of content that has been artificially generated or manipulated, including through disinformation (European Parliament, 2024). Additionally, the AI Act[4] stipulates that deployers, who use an AI system to generate "deepfakes," should also clearly and distinguishably disclose that the content has been artificially created or manipulated by Labeling the AI output. The compliance with this transparency obligation should not be interpreted as indicating that the use of the system or its output impedes the right to freedom of expression and the right to freedom of the arts and sciences (European Parliament, 2024). The requirement to label content (p. 136 of the AI Act) generated by AI systems is without prejudice to the obligation in Article 16 (Peukert, 2024) of Regulation (EU) 2022/2065 (European Union law, 2024) - providers of hosting services shall process any notices that they receive under the mechanisms to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content.

These norms will not be binding for Ukraine and VLOPs' activities in the country until Ukraine becomes an EU member. Thus, the norms should be reflected in Ukrainian regulations on digital services and AI.

Another example of non-legislative regulation in the EU is the Code of Practice on Disinformation of the EU. The Code of Practice was initially signed by Facebook, Google as well as Twitter, Mozilla, advertisers, and parts of the advertising industry, Microsoft and TikTok (European Commission, 2024). In the Code, the signatories recognized "the fundamental right to freedom of expression and to an open Internet, and the delicate balance which any efforts to limit the spread and impact of otherwise lawful content must strike" (European Commission, 2024). Special attention in the Code is given

1 p. 110.
2 p. 60 of the AI Act.
3 p. 120.
4 p. 134.

to the case law of The Court of Justice of the European Union (CJEU) on the proportionality of measures designed to limit access to and circulation of harmful content (European Commission, 2024).

It is worth mentioning that Russia has been expelled from the European Court of Human Rights (ECHR) so is no longer a state party to rulings on HR violations by that court, including on Article 10. However, in the recent ECHR Judgment dated 22 October 2024 in the case of Kobaliya and others v. Russia the Court held that there had been violations of the right to freedom of expression, the legislative framework had become considerably more restrictive since 2012, and had moved even further from Convention (European Convention on Human Right) standards" (European Court of Human Rights, 2024). As a result, such judgments of ECHR will not be currently legally executed in Russia. These make the ECHR mechanism not feasible for counter-disinformation efforts.

Contrary, the CJEU judgments could be a more efficient tool to counter Russian disinformation campaigns based upon the EU's regime imposing restrictive measures on Russian individuals and entities due to Russia's aggression against Ukraine (Pingen and Wahl, 2024).

Finally, the Code of Practice principles lay out two parts to consider when developing a legal framework for counteracting AI-powered disinformation: (1) no authority for governments to compel content moderation; (2) the content moderation could not be executed only on the basis that messages are thought to be "false."

3.2 Combating FIMI and counter AI-powered disinformation

AI development and accessibility have generated a lot of discussion, especially when it comes to how they could be abused for malevolent objectives in disinformation and FIMI. Actors in the FIMI rapidly started experimenting with these new tools to produce simulated media (EEAS, 2024a). The FIMI concept was developed to intercept various tactics used to manipulate society and protect the information space (EEAS, 2024b). Substantial disinformation efforts aiming at undermining EU members were uncovered following the start of Russian aggression in Ukraine in 2014, marking the first substantial moves in the FIMI approach. Officially implemented as part of EU policy, the specific FIMI effort was part of the European Action Plan Against Disinformation (European Union, 2018), which was enacted in December 2018. This action plan called for the establishment of a quick alert system to facilitate coordination amongst EU member states and the development of an operational task force, known as the East StratCom Task Force, to battle disinformation.

FIMI is defined by the European External Action Service (EEAS) as a pattern of behavior that jeopardizes or may have a detrimental effect on political structures, procedures, and ideals. These kinds of actions are planned, deliberate, and manipulative. While the main objectives of the EU are to safeguard the information space from harmful external influences, ensure transparency and honesty in information exchange, and strengthen democratic institutions by enhancing resilience to disinformation, the primary goal of this behavior is to influence public opinion, undermine democratic processes, and destabilize society (FIMI-ISAC, 2024).

Through collaboration with NGOs, academic institutions, and media from Africa and Latin America, the European Union and the United States are strengthening their preparedness for FIMI. To establish a multilateral community with the goal of enhancing collaboration in response to FIMI, they host workshops. To gain a deeper understanding of the disinformation strategies and narratives that are used in these areas, as well as the capacity of local stakeholders to counter them, they also gather data from fact-checking networks. Training in digital competency, media literacy, and development funding channels all improve support for capacity-building. It is also highlighted that maintaining a free and diverse media environment is crucial for countering disinformation and other FIMI (EEAS, 2024c).

Combating FIMI is essential to preserving national security, upholding democratic processes, and guaranteeing social cohesion within the EU, US, and other countries including Ukraine.

3.3 US legislation and regulation

As mentioned above the VLOPs' terms of service are developed based upon the legal system of the country of origin, mostly the US one. It is worth paying attention to the US regulations in order to figure out the principles used by VLOPs while addressing AI-powered disinformation.

The US legislation on AI-powered disinformation, which is important for the Ukrainian government to address Russian disinformation domestically and abroad, was drafted in the form of The Deepfakes Accountability Act introduced in June 2019 to combat the spread of disinformation through restrictions on deep-fake video alteration technology (Clarke, 2019). Additionally, the United States has introduced legal regulation with the Algorithmic Accountability Act (Wyden, 2022).

On the other hand, proposals for legislative ways to address disinformation in SM networks that give rise to national security concerns could contribute to the ongoing debate in the US and worldwide on the matter of how and by what authority SM networks could be regulated. Taking into account the evolving moderation of online expression, the gap in legal and regulatory terms regarding VLOPs' responsibility affects the national interests of other democratic states. The complexity of the problem derives also from the evolving opportunities for disinformation spurred by technology – for instance, mentioned above "deep-fakes" produced with the use of AI.

However, the US Constitution and case law have not been particularly consistent in the application and interpretation of freedom of speech restrictions. The US Constitution, First Amendment only applies to laws enacted by Congress and to local, state, or federal government agencies, but not to the actions of private VLOPs. Thus, the responsibility of VLOPs regarding freedom of speech and counter disinformation dissemination activities are defined predominantly by corporate policies. The US and other democracies' legal approaches to VLOPs were thus far based upon self-regulation. However, new regulations are currently evolving in AI and counter disinformation domain in the US and EU.

The US Deepfakes Accountability Act 2023 established the "Deepfakes Task Force" particularly to advance efforts of the US Government to combat the national security implications of deepfakes (Clarke, 2019).

A historic US Executive Order on Safe, Secure, and Trustworthy AI was also issued in late 2023, requiring systems that are far more sophisticated than those in use today to submit reports (The White House, 2023). Additionally, in the US NIST developed the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile with the preventative action to identify potential content provenance risks and harms in GAI, such as disinformation, deepfakes (NIST, 2024).

However, the US legislation on counter disinformation is developing the freedom of speech and freedom of VLOPs' entrepreneurship principles without proper consideration of the national security concerns. For instance, the Disinformation Governance Board Prohibition Act 2023 terminated the Disinformation Governance Board of the Department of Homeland Security (Bice, 2023). Additionally, the Free Speech Protection Act 2023 prohibits federal employees and contractors from directing online platforms to censor any speech that is protected by the First Amendment to the Constitution of the US (Paul, 2023).

As a result of a lack of bilateral US-Ukraine governmental collaboration tools, the Ukrainian government needs to develop legal and operational mechanisms to directly communicate with the VLOPs, based in the US, to counter Russian AI-powered disinformation. In this context, the principles of AI self-regulation, voluntarily committed and introduced in 2023 by several AI developers in the US (The white House, 2024), need to be particularly considered.

3.4 Ukraine's regulation on countering disinformation and AI

To understand the legal prerequisites for countering AI-powered disinformation in Ukraine there is a need to briefly analyze the country's relevant legislation.

Article 43 of the Ukrainian Constitution recognizes article 19 of ICCPR providing: "the exercise of… rights (to freedom of thought and speech) may be restricted by law in the interests of national security…" (Refworld, 2024b) The Ukrainian Law on Information defines "completeness and accuracy of information" (Bulletin of the Verkhovna Rada, 1992) as one of the basic principles of informational relationships. But there is no answer on the criteria of such information, and consequently no clear legal background for a definition of disinformation. Ukraine needs to establish this definition of disinformation in its national legislation. This is even more important considering the harm to the Ukrainian nationals caused by Russian disinformation campaigns. Executing the right to defense from Russian armed aggression Ukraine has additional justification on the basis of national security to counteract the adversary's disinformation operations. These included, for example, the May 2017 banning of several Russian websites such as VKontakte, Odnoklassniki, Yandex, and Mail.ru.

The first legal concern for Ukraine was to establish and enhance governmental authorities in charge of addressing disinformation. The necessity for the Ukrainian government to guard against foreign meddling in SM is obvious, but it could still potentially impinge upon freedom of expression and result in direct censorship. Another action the Ukrainian government made to counteract Russian propaganda and disinformation was the creation of the appropriate governmental bodies (e.g., Ministry of Culture and Strategic Communications). In March 2021, the Center for Strategic Communications and Information Security was established with the main objective of joint development of mechanisms to counter disinformation, together with international partners. Further, in line with the National Security Strategy of Ukraine, the CCD was established. The primary goal of this step was to ensure effective countering of propaganda and destructive disinformation campaigns. In May 2021 the President of Ukraine signed the regulation establishing the CCD (President of Ukraine, 2024a). The framework of this document establishes that the CCD is tasked with identifying and countering disinformation, propaganda, and destructive informational influence efforts and campaigns, as well as preventing attempts to manipulate public opinion (President of Ukraine, 2024b).

With regard to the existing Ukrainian legislative experience in the field of countering disinformation, Ukraine is gradually moving toward developing its own approach to countering the ever-growing information threats. Thus, in recent years, the Ukrainian legal framework has been supplemented by the Laws of Ukraine "On Cloud Services" and "On Stimulating the Development of the Digital Economy in Ukraine," as well as the Law "On Media."

At the same time, it should be emphasized that Ukraine has legislation in place to counter disinformation that provides guarantees of judicial protection and civil rights, among other things: For instance, Article 32 of the Constitution of Ukraine states: "Everyone is guaranteed judicial protection of the right to refute false information about himself or herself and members of his or her family and the right to demand the withdrawal of any information, as well as the right to compensation for material and moral damage caused by the collection, storage, use and dissemination of such false information" (Refworld, 2024b). Further, the Law on Information, Article 278 of the Civil Code of Ukraine, attempts to balance freedom of speech with the protection of legitimate interests, rights, and freedoms of individuals and legal entities. The Criminal Code of Ukraine (CCU) (Legislation of Ukraine, 2011) provides for criminal liability for certain offenses related to destructive information influence, in particular: Article 259 of the CCU ("Knowingly False Reporting of a Threat to the Safety of Citizens, Destruction or Damage to Property") provides for liability for knowingly false reporting of preparations for an explosion, arson, or other actions threatening the death of people or other severe consequences; Article 436 of the CCU ("Propaganda of War") provides for liability for public calls for aggressive war or for initiating a military conflict, as well as for the production of materials with calls to commit such actions for the purpose of their dissemination or distributing such materials. However, none of these legal acts provide for the liability of VLOPs and search services for the dissemination of disinformation, particularly with the use of AI.

In August 2023, the Parliament of Ukraine Verkhovna Rada adopted the "European Integration" law "On Digital Content and Digital Services," which introduced the terms "digital content" and "digital service." This law is aimed at protecting consumer rights when purchasing and using digital content or services. However, this law does not include a human rights part, as the European Digital Services Act (DSA) does. This opens up opportunities for Ukraine to define the legal relationship between VLOPs and users in the context of countering AI-powered disinformation. It should also be emphasized that as part of the implementation of measures to synchronize and harmonize Ukrainian legislation with that of the EU, the Ministry of

Digital Transformation and the Verkhovna Rada Committee on Humanitarian and Information Policy are taking active measures to implement the DSA, in particular by amending the Law of Ukraine "On Media."

Specifically, the DSA envisages new responsibilities for platforms and the empowerment of users, which, among other things, will include the following measures for:

- Countering illegal content, goods and services: Online platforms should provide users with the ability to flag illegal content, including goods and services. Moreover, platforms should cooperate with "trusted flaggers," specialized organizations whose alerts should be prioritized by the platforms.
- Protecting minors: including a complete ban on targeting minors with ads based on profiling or their personal data.
- Providing users with access to a complaint mechanism to appeal decisions on content moderation.
- Publishing a report on content moderation procedures at least once a year.
- Providing users with clear terms and conditions, including the basic parameters on which their content recommendation systems operate.
- Appointing a contact person for government authorities and users.

The implementation of the DSA in Ukraine will bring significant benefits in line with the countermeasures against AI-powered disinformation. First, it will increase the transparency of online platforms in Ukraine by forcing them to be open about their content moderation algorithms and procedures, providing users with information about the reasons for content removal or account blocking, and publishing annual reports. Second, it will strengthen the protection of users' rights by providing them with the opportunity to appeal content moderation decisions through special complaint mechanisms, which will help protect the rights to freedom of speech and personal information. Third, the law will help fight illegal content by allowing users to flag illegal content and cooperate with "trusted flaggers." The Concept on development of AI in Ukraine was approved by order of the Cabinet of Ministers of Ukraine on December 2, 2020 No. 1556 (The Parliament of Ukraine, 2020), however, there is no law on AI yet in Ukraine.

Finally, it is worth mentioning that Ukraine is actively working to put the FIMI standard into practice right now. The European External Action Service EEAS trained the CCD of the National Security and Defense Council of Ukraine in October 2023, working with the EU Advisory Mission in Ukraine (EUAM Ukraine) to introduce them to the FIMI approach and evaluate the suitability and effects of implementing FIMI for the CCD. The CCD receives from EEAS a complimentary dedicated instance of Open Cyber Threat Intelligence (OpenCTI), a knowledge management and sharing platform for FIMI and cyberspace.

Summing up, AI-powered disinformation campaigns undermine democratic processes, but is it enough to apply the freedom of speech exemptions based on national security concerns? Additionally, what could and should be the legal mechanism to clearly define national interests on a case-by-case basis? The answers are not obvious because of the nexus of domestic and international issues involved and the differences within legal systems. In the current circumstances, the Ukrainian government had legitimate grounds to proceed with banning AI-powered disinformation on VLOPs' platforms based on national security concerns. It is already recognized worldwide that Russia violated international law, and Ukraine had a right to impose anti-disinformation measures as a proportionate self-defense in line with international and domestic law.

4 Discussion: law applicability in AI and counter disinformation nexus to address national security concerns

4.1 Self-regulation or binding legal regulations for VLOPs

There is a complex legal issue centering on the responsibility of VLOPs, with regard to freedom of speech and the laws applicable to certain disinformation dissemination activities. As described above, Ukraine is dependent on corporate policies defining the responsibility of the US-based VLOPs regarding freedom of speech and counter disinformation, including AI-powered.

Despite professing commitment to free speech, the main objective of these companies is profit. The more customer attention VLOPs attract, the more advertising revenue is gained. Provided that disinformation is not defined as illegal and tends to spread further and faster than verified information, VLOPs can be potentially incentivized to engage in its dissemination. VLOPs faced the risks of "wasting" time, finance, and human resources on addressing AI-powered disinformation by monitoring their networks, detecting fake news and even losing their users if the moderation gave rise to public debate; thus, VLOPs used to be reluctant to identify perpetrators of disinformation.

Interestingly, according to the US Communications and Technology Subcommittee and the Consumer Protection and Commerce Subcommittee (House Committee on Energy and Commerce, 2024), the industry self-regulation has failed. This opinion is seconded by scientists, who believe that self-governance is not able to consistently endure the pressure of financial incentives. Assuming that these incentives will always be in line with the public interest is insufficient given AI's huge potential for both positive and negative effects. Governments need to start creating efficient regulatory frameworks right away if they want the development of AI to benefit everyone (Toner and Mccauley, 2024). Let us have a look at some VLOPs' internal policies and trends to counter AI-powered disinformation. VLOPs under pressure or in collaboration with governments, predominantly the US one, started to develop detection and suspension initiatives, including those relying on artificial intelligence, aimed at bots and botnets, as well users exposed to mis- and disinformation, reinforcing the visibility of reliable content produced by trustworthy media and fact-checking sources, and vice versa reducing visibility (Santa Clara University, 2024) or suspension of sites' disinformation content.

The creation of the Facebook (Meta) Oversight Board was, for instance, a positive step toward setting principles and rules for the VLOP content moderation. However, with no binding law regarding counter-disinformation, the Oversight Board can only solve its flagged concerns based on the Code of Conduct, which does not provide a clause on disinformation, particularly on

AI-powered content. Facebook's (Meta's) policy criteria of importance to public discourse and the number of individuals impacted is vital from the Ukrainian counter-disinformation perspective, such as in the case of the Kremlin-linked TV channels ban, where the need to prohibit broadcasting on SM platforms presented (Dickinson, 2021). Google legal policies (for instance, YouTube's) (Google, 2024) stipulate the following legal issues to file a complaint: trademark, counterfeit, defamation, stored music policy, other legal issues and complaints (YouTube, 2024). There is no exact and clear way to counter AI-powered disinformation using legal mechanisms. Even TikTok announces its initiatives to improve platform transparency and prevent covert influence campaigns. The platform claims to have discovered and destroyed networks involved in coordinated acts of inauthentic behavior (Tiktok, 2024).

However, we should agree with Rebecca Hamilton's opinion (Hamilton, 2021) that, "as a strictly legal matter, there is no reason for the platforms to have developed the elaborate content-moderation systems they currently run." Another complicated legal issue is that AI developers have motivations that are not in line with the interests of the general population. Developers will probably be pushed by financial incentives to underinvest in safety, which would be especially worrying if frontier AI systems result in significant negative externalities. This motivation mismatch indicates that there is also a need for strict supervision of AI developers. Thus, the only consistent solution at the national and/or international level would be to enact enforcement regulations covering VLOPs' operations in addressing AI-powered disinformation.

The case of Ukraine shows that rapid action at times has to be taken, and we want to show some examples of how this may be possible on the legal level. Of course, such rapid action is possible in the unprecedented circumstances of limited freedom of martial law. The legal concern is the extent to which governmental authority respects freedom of speech, privacy, and rule of law principles while addressing AI-powered disinformation. National governments should not be the only ones in charge of addressing AI-powered disinformation. Corporations should not be in charge of self-regulation either Marsden et al. (2020) propose co-regulation when businesses create their own user regulations, either separately or together, which must then be authorized by democratically legitimate state legislatures or regulators, who also keep an eye on how well they work. Such an approach could be effective in the Ukrainian realm of law while defending from Russian aggression. Accepting the principle that regulatory policies may be more reversible in AI than in other environments (Carpenter, 2024), we propose a "functional approach" (see Table 1), based upon the analysis of actions required for countering AI-powered disinformation: prevention, detection, and response to such campaigns.

4.2 Prevent, detect, and respond to AI-powered disinformation

From Ukraine's perspective, the Law on Countering Disinformation could be justified as an emergency measure against internationally recognized Russian military aggression combined with hybrid warfare. However, this law should nonetheless be in line with international law and recognized principles of freedom of speech. We propose the classification of Ukrainian authorities' powers with a set of preventative, detective, and responsive activities to address AI-enabled disinformation.

TABLE 1 Responsibility of stakeholders in counter AI-powered disinformation activities.

Actions\Stakeholders	State	VLOPs	Civil society organizations/ traditional media/ academia	Citizen(s)
Prevention				
Development of reliable news network	Support (S) 3	S2 (*number correlates with the level of involvement from 1 – highest to 3 lowest*)	L (*Leading stakeholder*)	S1
Raising awareness	L	S2	S1	S3
Providing mechanism for raising concern about national interests	L	S3	S1	S2
Facilitation of information-sharing platform	L/S1	S2	L/S1	S3
Detection				
Development of/enhancing algorithmic criteria for early detection of disinformation	S2	L	S1	S3
Fact/source-checking	S2	S3	L	S1
Response				
Strategic silence	L	S1	S2	S3
Strategic communication	L	S3	S1	S2
Sanctions and other economic and diplomatic measures	L	S3	S1	S2
Cyber information operations	L	S1	S3	S2
Flagging and dispelling	S3	L	S1	S2

4.2.1 Preventative actions

The development of a reliable news network is one of the first steps to take to prevent AI-powered disinformation. Our opinion is that democratic states should play a less active role in this activity than civil society organizations/traditional media, citizens, and VLOPs because of freedom of speech constitutional guarantees. Exceptions could be made only on the precondition of martial law limits. Strategies for raising awareness about AI-powered disinformation threats to national security should be among the measures for the state to undertake, in order to strengthen society's compliance. As part of regulating and countering AI-powered disinformation, states should provide mechanisms to raise national interest concerns based upon academic research, while taking into consideration intelligence community analysis.

For Ukraine, such a procedure could involve CCD's proposals to NSDC based on Security and Defense agencies' analyses and civil society organizations/traditional media/academia inputs. The vital point of this mechanism is the implementation of NSDC decisions by VLOPs. Public-private cooperation in counter-disinformation requires knowledge-sharing between governments, VLOPs, and other stakeholders. An experience-based, lessons-learned platform, to share knowledge of adversaries' methods and techniques, etc. can be developed in Ukraine, based on the example of disinfocloud.com, an online platform provided by the US Global Engagement Center to connect with relevant stakeholders (Global Engagement Center, 2024).

In the EU there is a different approach: an independent, non-profit organization focused on tackling sophisticated disinformation campaigns targeting the EU, its member states, and core institutions – Disinfo Lab (EU DisinfoLab, 2024). An important action in preventive measures of AI-powered disinformation is the ongoing education and awareness-raising agenda among the actors involved in combating disinformation. This includes a set of measures aimed at raising the level of media literacy and information hygiene among the population.

In an environment where the information space is filled with a large amount of destructive content, the ability to critically evaluate information becomes vital. For example, teaching citizens to distinguish facts from opinions or propaganda helps protect them from disinformation, including from that generated by AI. An important aspect of media literacy is also understanding the algorithmic mechanisms that govern the presentation of content on SM and news platforms, which allows for a better understanding of why people see certain content. Educational activities to improve media literacy and information hygiene should be systematic and cover all age groups. This can be done through educational programs at schools and universities, training for adults, as well as through the media and social networks. Particular attention should be paid to the younger generation, who are active users of digital technologies and are particularly vulnerable to disinformation. Successful implementation of these measures will contribute to the creation of a more resilient society that can effectively resist destructive information influences.

The Russian war against Ukraine has shown that media literacy is not only an academic topic for discussion but also an important process of developing relevant skills that save health and life. In general, the promotion of media literacy in Ukraine is part of a broader strategy aimed at creating an informed society. Recognizing these threats, Ukrainian authority has been actively engaged in public awareness programs and campaigns and cooperation with civil society organizations to promote media literacy as a tool to strengthen the country's information resilience (Horban and Oliinyk, 2024).

Finally, scientists advocate for a global consensus on the ethical usage of GenAI and implementing cyber-wellness educational programs to enhance public awareness and resilience against disinformation (Shoaib et al., 2023).

4.2.2 Detection

The best counter AI-powered disinformation response in SM, arguably, are algorithmic approaches to detecting disinformation before it becomes shareable. These actions require clear legal regulation of VLOPs responsibility to detect AI-powered disinformation. AI-powered disinformation campaigns can rarely be detected at the early stages, for instance, adversaries' research on the audience or narratives and fake news preparation including making the AI-powered disinformation credible. Such activities can only be detected by clandestine operations of the intelligence communities.

However, VLOPs actually have the technical capabilities "*to detect mis- and disinformation in real time*" (Bharat, 2017). The basic criteria to qualify some activity as AI-powered disinformation could be if the activity: developed or disseminated by AI system; contains deceptive elements; has the intention to harm; is disruptive; constitutes interference" (Pamment et al., 2024). Such criteria must be available to the public, if used by detection tools.

In terms of impact, the detection of AI-powered disinformation could be made using AI, before fake news dissemination occurs in SM. VLOPs already use their AI-based products to provide feedback to commenters about potential perceived toxicity of content in real-time (for instance Jigsaw's Perspective and Tune). This is a valuable tool for individuals, which allows readers to choose the level of toxicity they will see in comments across the internet (Jigsaw, 2024). Scientists like Smith et al. (2021) propose an end-to-end system to perform narrative detection, hostile influence operations account classification, network discovery, and estimation of hostile influence operations causal impact; as well as a method for detection and quantification of causal influence on a social network. Such results could be used by the Ukrainian authority to detect AI-powered disinformation.

The technical approach proposed by Nitzberg and Zysman (2022) for enabling AI to slow down the amplification of disinformation messages by the employment of time-limitation features for sharing suspected messages, could be efficient at a post-detection stage. If the message is not confirmed to contain fake elements, it could be disseminated at the usual pace; otherwise, it should be flagged or dispelled.

The next action to counter AI-powered disinformation is fact-checking. The authors propose that a fact-checking mechanism be used as a detection activity – before dissemination of what is suspected by AI to be false news. At present, fact-checking occurs after the incriminated fake news has been disseminated. This approach, however, is not sufficient. False information is diffused and has a "*continued influence effect*" (Lewandowsky et al., 2012). Thus, it is vital to develop a proactive counter-AI-powered disinformation detection mechanism. The fact-checking tools developed by VLOPs (mentioned above) and Ukrainian projects like StopFake (2024) and VoxUkraine

(2024) etc., could all contribute to the Poynter Institute's international network (IFCN Code of Principles, 2024).

There were and still are certain factors that could influence early detection of fake news: for instance, a debate about using encryption, particularly in VLOPs (National Academies of Sciences, Engineering, and Medicine, 2018). These debates seriously challenge counter AI-powered disinformation measures at the detection stage, from both a legal and technical perspective.

It may also be necessary for developers to employ identifiers that permit the identification of content produced by AI. One of the ways to counter the above-mentioned is to create special AI-based software that will mark the content in which the visual or audio part has been interfered. This will alert the user about the possible danger of using the "real-time" deepfake. However, the development of such software requires a long time and significant resources, which makes it impossible to counter this destructive influence. Therefore, the Ukrainian experience proposes to counteract "real-time" deepfake calls by carefully verifying the facts of the planned online meeting, staying in contact only through official means of communication, using different communication channels when organizing a video call, and following basic rules of cyber hygiene.

4.2.3 Response actions

Even though attribution in AI-powered disinformation efforts might be challenging, it's crucial to coordinate attribution and response when sufficient evidence is available and to publicly denounce those who spread false information (Kertysova, 2018). The response actions against AI-powered disinformation are dependent on the attribution of hostile influence campaigns, which is difficult. Response actions are also contingent on jurisdiction, which defines the mechanisms for decision-making as well as the status of data in transit. VLOPs, for instance, can change the data transactions from one jurisdiction to another using their servers' locations and business process requirements. The same could be done by perpetrators to hide the tracks of disinformation dissemination. In consequence, this would severely complicate the attribution of AI-powered disinformation.

Ukrainian experience shows that one of the most important tools for responding to destructive information influence is the development of positive strategic narratives that help build society's resilience to disinformation, particularly that powered by AI. These narratives strengthen trust in official sources and create a positive image of the state in the international arena. The development of such narratives involves the dissemination of new and reliable materials with the involvement of experts from academia, civil society, foreign language experts, media representatives, and partners from other democratic countries. This ensures a high level of diversity and reliability of information. In general, positive strategic narratives should be based on real achievements and events that build trust in information sources. They should be understandable and relatable to the audience, taking into account the values and interests of the latter. This case shows that positive strategic narratives are a powerful tool for countering AI-powered disinformation and strengthening information security and society's resilience to external influences.

The same applies to the identification of negative (hostile) strategic narratives, which is crucial for countering AI-powered disinformation. Identifying dangerous messages aimed at discrediting state institutions, undermining trust in official sources, and creating panic among the population is a top priority. Analyzing the purpose and context of hostile narratives allows us to understand the goals behind the messages. Identifying the tactics and methods used in hostile narratives, such as intimidation, divergence, and fake news enables to develop of strategic countermeasures to neutralize their impact. It is worth noting that developing positive strategic narratives and identifying negative strategic narratives are the primary countermeasures against the influence of destructive information, as through such actions it becomes possible to identify global directions for countering AI-powered disinformation and the main steps toward it.

Another way to respond to AI-powered disinformation – disregard or strategic silence – should be considered in government decision-making taking into account that public opinion would have a tendency to forget quickly. Strategic silence could be used when the risk of danger for national security narratives to be perceived and believed by the population is low. However, this method carries significant potential risks. The main danger lies in the possible incorrect determination of the threat level. If the AI-powered disinformation that is decided to be ignored has a high level of disruptive impact, ignoring it can have serious consequences. For example, it may increase the spread of harmful narratives that can negatively affect public opinion, increase distrust of state institutions, or even cause panic. Therefore, the decision to use strategic silence should be made based on a thorough analysis of the potential impact of AI-powered disinformation. It is important to take into account not only the current state of public opinion but also the potential long-term consequences that may arise from underestimating the threat.

Strategic communication as a way to respond to AI-powered disinformation aims to provide and disseminate new and truthful content; this approach requires time, resources and a systemic framework. The use of humor as a part of responding to disinformation will also help to increase the dissemination of counter disinformation messages on SM platforms. Sanctions and other economic and diplomatic measures are additional legal tools to respond to AI-powered disinformation. One example of a sanction is the US legislation mandating the sale of TikTok based on concerns over disinformation and foreign propaganda (Fung, 2024).

Informational sanctions (flagging or blocking SM accounts) is an approach proposed by the authors, for further consideration and possible use against entities and individuals involved in AI-powered disinformation. In the context of the implementation of the information sanctions mechanism in Ukraine, it is necessary to emphasize a number of important tasks of the CCD at the NDC, including analysis and monitoring of events and phenomena in the country's information space, assessment of the state of information security and analysis of Ukraine's presence in the global information space. One of the key aspects of the CCD's activities is the identification and study of current and predicted threats to Ukraine's information security.

Rapid identification of the main actors generating AI-powered disinformation is crucial for an effective countering of information threats. The CCD closely cooperates with state authorities, law enforcement, and intelligence agencies, including foreign ones, to provide selected and analyzed data on key actors generating

AI-powered disinformation. This data is passed on to the appropriate authorities for imposing sanctions and decision-making.

This innovative model of work of the CCD provides a comprehensive approach to countering AI-powered disinformation. It includes not only the identification and analysis of threats but also active cooperation with various national and international organizations, which allows them to respond quickly to changes in the information environment and effectively counter AI-powered disinformation campaigns. In particular, the CCD uses modern technologies to monitor the information space, analyze large amounts of data, and predict potential threats. This includes the use of AI algorithms to automatically detect anomalies in media content. The results of such analysis allow us to accurately identify sources of AI-powered disinformation and assess their impact on society.

Response in the form of cyber information operations is conducted covertly. The malicious use of general-purpose AI for deception and public opinion manipulation is a further topic of concern. AI-powered disinformation can be produced by adversaries and spread more easily even with the aim of influencing political processes.

Response actions include flagging and dispelling fake messages in SM. However, flagging or labeling fake information as "disputed" is not successful because it causes more sharing of the flagged content, and merely labeling information as fake does not lead to a reduction in its spread (Smith, 2017). One of the proposed ways to address the issue of deepfakes is to create a digital watermarking system that can verify the authenticity of media content (Thumos, 2024). Watermarking, which employs an invisible signature to identify digital content as coming from or being updated by AI, is one recommended technique for spotting disinformation (Christ et al., 2024). Although they are helpful, technical countermeasures like content watermarking are typically vulnerable to reasonably skilled offenders (AI Safety Institute, 2024).

The extent of governmental authority to counteract AI-powered disinformation with respect to freedom of speech, privacy, and rule of law principles is shown in Table 1.

Authors assign each action mentioned in the table to stakeholders based upon the following considerations: (1) the state cannot exercise influence on the development of a reliable news network and the fact/source-checking process, apart from the official governmental platform; (2) developing and enhancing algorithmic criteria for early detection of AI-powered disinformation, as well as flagging and dispelling it, are the responsibility of VLOPs due to their technical capacity; (3) the state authority should be able to choose the proper response to AI-powered disinformation in order to counter it (excluding flagging and dispelling).

5 Conclusion

AI-powered disinformation is becoming increasingly present in our lives and addressing it should be high on the agenda of national governments and interstate entities. Specifically, the legal means must be adjusted, based on detailed analyses of counter disinformation ecosystem, international and national legislation, as well as emerging regulations on AI systems. The European Media Freedom Act, the Future of EU Digital Policy, the EU Code of Practice on Disinformation, and the European AI Act already contain some norms for the regulation of AI-powered disinformation. When it comes to the work of very large online platforms (VLOPs), their internal counter-disinformation policies are largely oriented on the US liberal legislation in counter-disinformation, as most VLOPs are headquartered in the U.S.

Amid these realities, the transformations taking place in Ukraine present a case of particular interest. The country's government is actively harmonizing its legislation with the EU binding legislation and regulations in AI- and counter-disinformation measures. Ukrainian Law on counter-disinformation measures, developed as an emergency response to internationally recognized Russian military aggression and hybrid warfare tactics, underscores the crucial need to align even emergency measures with international law and principles of free speech.

The authors proposed a set of preventative actions. These are developing reliable news networks, raising awareness, providing a mechanism for raising concerns about national interests, and facilitating information-sharing platforms. Detection actions are defined as developing/enhancing algorithmic criteria for early detection of disinformation, and fact/source-checking. Response actions are defined as strategic silence, strategic communication, sanctions and other economic and diplomatic measures, cyber information operations, and flagging and dispelling.

Author contributions

AM: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. SP: Data curation, Formal analysis, Writing – original draft, Writing – review & editing. AK: Data curation, Formal analysis, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, 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.

Publisher's note

References

AI Safety Institute. (2024). International scientific report on the safety of advanced AI. Available at: (https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai)

AI Safety Summit. (2023). The Bletchley declaration by countries attending the AI safety summit. Available at: (https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023)

AI Seoul summit. (2024). Frontier AI safety commitments. Available at: (https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024)

Anderljung, M., and Hazell, J.. (2023). Protecting society from AI misuse: when are restrictions on capabilities warranted? Available at: (http://arxiv.org/abs/2303.09377)

Artificial intelligence. (2023). AI Safety Summit: introduction. Available at: (https://www.gov.uk/government/publications/ai-safety-summit-introduction)

Bharat, K. How to detect fake news in real-time. (2017) Available at: (https://medium.com/newco/how-to-detect-fake-news-in-real-time-9fdae0197bfd)

Bice, S. I.. (2023). Disinformation governance board prohibition act. Available at: (https://www.congress.gov/bill/118th-congress/house-bill/4514/text)

Buiten, M. C. (2019). Towards intelligent regulation of artificial intelligence. *Eur. J. Risk Regul.* 10, 41–59. doi: 10.1017/err.2019.8

Bulletin of the Verkhovna Rada. (1992). Law of Ukraine on information. Available at: . (https://wipolex-resources-eu-central-1-358922420655.s3.amazonaws.com/edocs/lexdocs/laws/en/ua/ua055en.pdf)

Carpenter, D. (2024). Approval regulation for frontier artificial intelligence: Pitfalls, plausibility, optionality. Cambridge, MA: Harvard Kennedy School.

Cavaliere, P. (2022). "The truth in fake news: how disinformation laws are reframing the concepts of truth and accuracy on digital platforms" in In: European convention on human rights law review. eds. K. Dzehtsiarou and V. P. Tzevelekos (Boston, MA: Brill Nijhoff).

Center for AI Safety. (2024). Statement on AI risk. Available at: (https://www.safe.ai/work/statement-on-ai-risk)

Center for Countering Disinformation. (2024). Analytical report "information influence campaign in the African information space." Available at: (https://cpd.gov.ua/en/report/analytical-report-information-influence-campaign-in-the-africaninformation-space/)

Christ, M., Gunn, S., and Zamir, O.. Undetectable watermarks for language models. In: The thirty seventh annual conference on learning theory. Edmonton, Canada: PMLR; (2024).

Clarke, Y. D.. (2019). Deepfakes accountability act. Available at: (https://www.congress.gov/bill/116th-congress/house-bill/3230/text)

Council of the European Union. (2024a). The future of EU digital policy. Available at: . (https://data.consilium.europa.eu/doc/document/ST-9957-2024-INIT/en/pdf)

Council of the European Union. (2024b). Council conclusions on democratic resilience: Safeguarding electoral processes from foreign interference. Available at: (https://data.consilium.europa.eu/doc/document/ST-10119-2024-INIT/en/pdf)

Dickinson, P. (2021). Analysis: Ukraine bans kremlin-linked TV channels. Available at: (https://www.atlanticcouncil.org/blogs/ukrainealert/analysis-ukraine-bans-kremlin-linked-tv-channels/)

EEAS. (2024a). 2nd EEAS report on foreign information manipulation and interference threats. Available at: (https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en)

EEAS. (2024b). Beyond Disinformation - What is FIMI? Available at: (https://www.eeas.europa.eu/eeas/beyond-disinformation-what-fimi_en)

EEAS. (2024c). TTC Ministerial - FIMI. Available at: (https://www.eeas.europa.eu/sites/default/files/documents/2023/Annex%203%20-%20FIMI_29%20May.docx.pdf)

ENISA. (2024). Skills shortage and unpatched systems soar to high-ranking 2030 cyber threats. Available at: (https://www.enisa.europa.eu/news/skills-shortage-and-unpatched-systems-soar-to-high-ranking-2030-cyber-threats)

EU DisinfoLab. (2024). EU DisinfoLab. Available at: (https://www.disinfo.eu/)

European Commission. (2024). The code of practice on disinformation. Available at: . (https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation)

European Court of Human Rights. (2024). ECHR: Judgment concerning the Russian Federation. Available at: (https://www.echr.coe.int/w/judgment-concerning-the-russian-federation-16)

European Media Freedom Act. (2024). Framework for media services in the internal market and amending. Available at: (http://data.europa.eu/eli/reg/2024/1083/oj/eng)

European Parliament. (2024). Artificial intelligence act. Available at: (http://data.europa.eu/eli/reg/2024/1689/oj/eng)

European Union. (2018). Action plan against disinformation 52018JC0036. Available at: (https://eur-lex.europa.eu/legal-content/GA/TXT/?uri=CELEX:52018JC0036)

European Union law. (2024). Regulation - 2022/2065. Available at: (https://eur-lex.europa.eu/eli/reg/2022/2065/oj)

FIMI-ISAC. (2024). What is the mission of the FIMI-ISAC? Available at: (https://fimi-isac.org)

Fung, B.. (2024) Biden just signed a potential TikTok ban into law. Here's what happens next. Available at: (https://www.cnn.com/2024/04/23/tech/congress-tiktok-ban-what-next/index.html)

Funke, D., and Flamini, D.. (2024). A Guide to anti-misinformation actions around the world. Available at: (https://www.poynter.org/ifcn/anti-misinformation-actions/)

Gamito, M. C. (2023). The European media freedom act (EMFA) as meta-regulation. *Comput. Law Secur. Rev.* 48:105799. doi: 10.1016/j.clsr.2023.105799

Global Engagement Center. (2024). United States Department of State. Available at: (https://www.state.gov/bureaus-offices/under-secretary-for-public-diplomacy-and-public-affairs/global-engagement-center/)

Google. (2024). How Google fights disinformation. Available at: (https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf)

Hajli, N., Saeed, U., Tajvidi, M., and Shirazi, F. (2022). Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *Br. J. Manag.* 33, 1238–1253. doi: 10.1111/1467-8551.12554

Hamilton, R. J. (2021). Governing the global Public Square. *Harv. Int. Law J.* 62:117.

Heim, L., Anderljung, M., and Belfield, H.. (2024). Lawfare: to govern AI, we must govern compute. Available at: (https://www.lawfaremedia.org/article/to-govern-ai-we-must-govern-compute)

Horban, Y., and Oliinyk, O. (2024). Media literacy as a factor in protecting the information space from enemy disinformation in time of war. *Ukr. Inf. Space* 13, 194–205.

Horvitz, E. On the horizon: interactive and compositional Deepfakes. In: Proceedings of the 2022 international conference on multimodal interaction. Bengaluru, India: ACM; (2022).

House Committee on Energy and Commerce. (2024). The pressure is on big tech. Available at: (https://energycommerce.house.gov/posts/energycommerce.house.gov)

IFCN Code of Principles. (2024). Commit to transparency — sign up for the International Fact-Checking Network's code of principles. Available at: (https://www.ifcncodeofprinciples.poynter.org/)

Jigsaw. (2024). Machine learning can help reduce toxicity, improving online conversation. Available at: (https://jigsaw.google.com/the-current/toxicity/)

Kertysova, K. (2018). Artificial intelligence and disinformation: how AI changes the way disinformation is produced, disseminated, and can be countered. *Secur. Hum. Rights* 29, 55–81. doi: 10.1163/18750230-02901005

Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., et al. (2024). Responsible reporting for frontier AI development. Available at: (http://arxiv.org/abs/2404.02675)

Legislation of Ukraine. (2011). The Criminal Code of Ukraine. Available at: (https://zakon.rada.gov.ua/laws/show/2001-05#Text)

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: continued influence and successful Debiasing. *Psychol. Sci. Public Interest* 13, 106–131. doi: 10.1177/1529100612451018

Mahler, T.. (2021). Between risk management and proportionality: the risk-based approach in the EU's artificial intelligence act proposal. Available at: (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4001444)

Marsden, C., Meyer, T., and Brown, I. (2020). Platform values and democratic elections: how can the law regulate digital disinformation? *Comput. Law Secur. Rev.* 36:105373. doi: 10.1016/j.clsr.2019.105373

National Academies of Sciences, Engineering, and Medicine (2018). Decrypting the encryption debate: a framework for decision makers. Washington, D.C.: National Academies Press.

NATO. (2024). NATO: Washington summit declaration. Available at: (https://www.nato.int/cps/en/natohq/official_texts_227678.htm)

NIST (2024). Artificial intelligence risk management framework: generative artificial intelligence profile. Gaithersburg, MD, USA: NIST.

Nitzberg, M., and Zysman, J. (2022). Algorithms, data, and platforms: the diverse challenges of governing AI. *J. Eur. Public Policy* 29, 1753–1778. doi: 10.1080/13501763.2022.2096668

Osavul. (2024). AI-powered security against information threats. Available at: (https://www.osavul.cloud/)

Pamment, J., Nothhaft, H., and Fjällhed, A.. (2024). Countering information influence activities: the state of the art. Available at: (http://lup.lub.lu.se/record/825192b8-9274-4371-b33d-2b11baa5d5ae)

Paul, R. (2023). Free speech protection act. Available at: (https://www.congress.gov/bill/118th-congress/senate-bill/2425/text)

Peukert, A. (2024). The regulation of disinformation: a critical appraisal. *J. Media Law* 16, 1–7. doi: 10.1080/17577632.2024.2362485

Pingen, A., and Wahl, T.. (2024). General court judgments on anti-war sanctions. Available at: (https://eucrim.eu/news/general-court-judgments-on-anti-war-sanctions/)

President of Ukraine. (2024a). Presidential Decree №106/2021. Available at: (https://www.president.gov.ua/documents/1062021-37421)

President of Ukraine. (2024b). Presidential Decree №187/2021. Available at: (https://www.president.gov.ua/documents/1872021-38841)

Refworld. (2024a). General comment no. 34, article 19, freedoms of opinion and expression. Available at: (https://www.refworld.org/legal/general/hrc/2011/en/83764)

Refworld. (2024b). Constitution of Ukraine. Available at: (https://www.refworld.org/legal/legislation/natlegbod/1996/en/42875)

Roberts, H., Hine, E., Taddeo, M., and Floridi, L. (2024). Global AI governance: barriers and pathways forward. *Int. Aff.* 100, 1275–1286. doi: 10.1093/ia/iiae073

Rosen, B., Fang, K., and Shah, P.. (2023). Just security a right to spy? The legality and morality of espionage. Available at: (https://www.justsecurity.org/85486/a-right-to-spy-the-legality-and-morality-of-espionage/)

Rutube. (2024a). Prank c glavoy MID Latvii Krishyanisom Karinsh. Available at: (https://rutube.ru/video/f1b81b39c9e83d0b9afedc3b8d8583ad/)

Rutube. (2024b). Prank s Aleksandrom Kvasnevskim. Available at: (https://rutube.ru/video/2923db90d59af8f76faa9e3b86b0e975/)

Santa Clara University. (2024). The trust project helps readers identify reliable news. Available at: (https://www.scu.edu/news-and-events/press-releases/2017/nov-2017/the-trust-project-helps-readers-identify-reliable-news.html)

Scherer, M. U. (2015). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harv. J. L. Tech.* 29:353. doi: 10.2139/ssrn.2609777

Schuett, J. (2023). Defining the scope of AI regulations. *Law Innov. Technol.* 15, 60–82. doi: 10.1080/17579961.2023.2184135

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., and Zhao, J.. Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In: 2023 international conference on computer and applications (ICCA). Cairo, Egypt: IEEE; (2023).

Smith, J.. (2017) Designing against misinformation. Available at: (https://medium.com/designatmeta/designing-against-misinformation-e5846b3aa1e2)

Smith, S. T., Kao, E. K., Mackin, E. D., Shah, D. C., Simek, O., and Rubin, D. B. (2021). Automatic detection of influential actors in disinformation networks. *Proc. Natl. Acad. Sci.* 118:e2011216118. doi: 10.1073/pnas.2011216118

Spitale, G., Biller-Andorno, N., and Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Sci. Adv.* 9. doi: 10.1126/sciadv.adh1850

Stiglitz, J. E.. (2024). Foreign policy. Why big Tech's digital trade rules are harmful. Available at: (https://foreignpolicy.com/2024/04/04/big-tech-digital-trade-regulation/)

StopFake. (2024). StopFake. Available at: (https://www.stopfake.org)

The Parliament of Ukraine. (2020). The concept of development of AI in Ukraine. Available at: (https://zakon.rada.gov.ua/go/1556-2020-%D1%80)

The White House. (2023). Safe, secure, and trustworthy development and use of artificial intelligence. Available at: (https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence)

The white House. (2024). Fact sheet: Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. Available at: (https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/)

Thumos. (2024) Generative AI and deepfakes. How AI will create disinformation. Available at: (https://thumos.global/generative-ai-and-deepfakes/)

Tiktok. (2024). Countering influence operations. Available at: (https://www.tiktok.com/transparency/en/countering-influence-operations/)

Toner, H., and Mccauley, T.. (2024). AI firms mustn't govern themselves, say ex-members of OpenAI's board. Available at: (https://www.economist.com/by-invitation/2024/05/26/ai-firms-mustnt-govern-themselves-say-ex-members-of-openais-board)

Tregubov, V.. (2021). Ukraine raised the alarm over weaponized social media long before Trump's twitter ban. Available at: (https://www.atlanticcouncil.org/blogs/ukrainealert/ukraine-raised-the-alarm-over-weaponized-social-media-long-before-trumps-twitter-ban/)

Viljanen, M., and Parviainen, H. (2022). AI applications and regulation: mapping the regulatory strata. *Front. Comp. Sci.* 3:779957. doi: 10.3389/fcomp.2021.779957

VoxUkraine. (2024). Russia's deflections and denials about chemical weapons use in Ukraine. Available at: (https://voxukraine.org/en/category/voxcheck)

Wyden, R.. (2022). Algorithmic accountability act of 2022. Available at: (https://www.congress.gov/bill/117th-congress/senate-bill/3572/text)

Yampolskiy, R. V.. (2024). On monitorability of AI. AI Ethics. Available at: (https://link.springer.com/10.1007/s43681-024-00420-x)

YouTube. (2024). Legal policies. Available at: (https://support.google.com/youtube/topic/6154211?hl=en)

frontiers | Frontiers in Political Science

TYPE Original Research
PUBLISHED 24 January 2025
DOI 10.3389/fpos.2025.1434240

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Pilar Lacasa,
International University of La Rioja, Spain
Caner Çakı,
Ahi Evran University, Türkiye

*CORRESPONDENCE
Artem Zakharchenko
✉ artem.zakh@gmail.com

RECEIVED 17 May 2024
ACCEPTED 08 January 2025
PUBLISHED 24 January 2025

CITATION
Zakharchenko A (2025) Advantages of the connective strategic narrative during the Russian–Ukrainian war.
Front. Polit. Sci. 7:1434240.
doi: 10.3389/fpos.2025.1434240

Advantages of the connective strategic narrative during the Russian–Ukrainian war

Artem Zakharchenko[1,2]*

[1]Institute of Journalism, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, [2]NGO Communication Analysis Team – Ukraine, Kyiv, Ukraine

In this article, we expand the methodological approach to strategic narrative analysis based on the case of the contemporary Russian–Ukrainian war. Namely, we introduce the concept of a "connective strategic narrative." Such a narrative is not intentionally constructed by elites but created by the "affective public" on social media—emotionally tied social media users, according to Papacharissi's definition. We argue that the patriotic Ukrainian narrative about the war evolved in social media can be considered a connective strategic narrative that is more comprehensive than the "normal" strategic narrative shaped by authorities, while the pro-Russian social media war narrative is merely a reflection of the official strategic narrative. Based on social media data, we conducted a structural narrative analysis of both strategic narratives used in the ongoing war in Ukraine and deployed in the Ukrainian information space: the offensive pro-Russian narrative and the defensive Ukrainian narrative. The pattern for such analysis is based on Korostelina's framework for national narrative analysis. Our analysis emphasizes the key differences between these narratives and shows that the Russian one has crucial disadvantages that prevent it from successfully engaging the Ukrainian people. Instead, as it was developed with the significant participation of ordinary citizens, the Ukrainian strategic narrative had the total advantage in the struggle for the attention of Ukrainians at the beginning of the full-scale invasion.

KEYWORDS

strategic narrative, war communication, social media, affective public, Russian-Ukrainian war, propaganda

1 Introduction

Russia started full-scale aggression against Ukraine on 24 February 2022, after an eight-year "local" war in the East and South of Ukraine. The Kremlin had not achieved its goal of taking Kyiv by storm in three days; as of December 2024, this goal has not been achieved. One of the reasons for this is the significant resilience and will to fight demonstrated by the Ukrainian military and civilians, as noted by the media (Burns, 2022). This results from informational warfare as a part of the hybrid military confrontation, as described by military researchers (Caliskan, 2019). Therefore, the course of the invasion has shattered both the myth of Russian military might and the effectiveness of Russian propaganda and propagandists. This is possible when the offensive Russian strategic narrative spread in the Ukrainian information space was less effective than the defensive Ukrainian strategic narrative. However, the reason for this difference has not yet been clarified. At least, at the beginning of the conflict, both narratives had to establish their positions in the information space. The study period will be from 24 February to 4 April 2022. The final date was chosen as it was just after the substantial turning point in the attitude to the war, following the exposure of Russia's military crimes in Bucha.

Western military experts typically emphasize the high proactivity of Ukrainian soldiers and military units. They note (Hall, 2022) that the decision-making in the Ukrainian army is largely decentralized, which provides a crucial advantage on the battleground compared to the highly centralized Russian army. The same situation applies to the support of the army by Ukrainian volunteers. Communication in Ukrainian and social media is also decentralized and pluralistic, with many grassroots movements (Zakharchenko et al., 2019). In view of the above, we can hypothesize that the reason for the Ukrainian strategic narrative advantage can also be rooted in decentralization.

But how can it be decentralized? As an instrument of soft power, strategic narrative allows states to win the war. It is usually considered a story intentionally constructed by the authorities and told for their citizens and the enemy forces (Szostek, 2017a, p. 5). Therefore, it is created in a centralized manner.

To resolve this inconsistency, we apply the "affective public" concept introduced by Papacharissi (2015). We provide the concept of "connective strategic narrative" to show that in a time of social media communication, the actors of the strategic narrative are not limited to the authorities. Here, we consider strategic narrative as a story that motivates people to struggle.

This approach allows us to make a connection between the strategic narrative and the more ordinary social media narrative, which has been studied for many years as a comprehensive story about the world created on social media by different countries, generations, and more (Carmen et al., 2023). Therefore, we can hypothesize that the segment of this narrative that was formed in Ukraine at the beginning of the full-scale invasion plays the role of a strategic narrative. In the example of the Ukrainian war, we will answer whether the narrative created on social media can serve as strategic.

> RQ1. Is the social media narrative about the war in Ukraine appeared in the patriotic, pro-Ukrainian segment of social media and served as the strategic narrative of the Ukrainian side of this war during the first months of this war?

Pro-Russian social media content is also present in the Ukrainian media space and used by Russia to influence the Ukrainian audience (Katerynchuk, 2017). It is also expected to be a source for the Russian strategic narrative study, considering the strong centralization of the Russian propaganda machine and its control over communication in social media (Horbyk et al., 2023)—namely, Russia banned social media platforms that were not controlled by the state on its own territory as well as on the occupied territories. Therefore, the Russian social media narrative about the war cannot be anything other than just a mirror of the official narrative. But we have to check this before move forward.

> RQ2. Is the pro-Russian social media narrative in the Ukrainian information space about the war in Ukraine just a mirror of the official Russian strategic narrative presented by official Russian speakers?

Positive answer on RQ1 and RQ2 will enable us to compare the Ukrainian and Russian strategic narratives based on social media data.

> RQ3. What are the differences between the offensive Russian and defensive Ukrainian social media narratives about the war in its initial stage?

Determining why the Russian strategic narrative was unsuccessful in Ukraine is also important. This can be achieved by examining the requirements articulated by previous researchers who have worked on this topic:

> RQ4. Are there any significant problems in the Russian strategic narrative in Ukraine according to the requirements of the "good strategic narratives" articulated by scholars?

To answer all these questions, we develop a structural approach for strategic narrative analysis. It is common to think that this narrative is a kind of "artwork" that must meet a list of requirements—so there is less information about its construction. But, our approach allows us to show how to construct these narratives rather than create them. Within the context of social media, we consider narrative as a dynamic structure rather than a completed artwork, so it may differ significantly at various stages of the war.

2 Literature review

2.1 Identity and action narratives

Lyotard's «grand narratives» despite postmodern criticisms (Lyotard, 1984), still have a significant role as stories that bring people together and motivate them to take common action. In an age of mediated reality, these narratives are often shaped in a media communication space (Kaun and Fast, 2013).

There are two basic types of these narratives. The first is the identity narrative, which describes different peoples' identities, origins, rules, and goals. The most well-studied is the national narrative (Snyder, 2004). However, there are also other identity narratives, such as religious narratives, which are among the most widespread in human culture (Pihlaja, 2011), and political ideology narratives (Kaye, 2016). Gender, professional, subculture, and other narratives can also be suggested.

The structure of the national narrative is well-described by Korostelina (2014, p. 23–51). This narrative creates a national identity and legitimizes power in the country. Typically, such a narrative consists of three parts:

- Binary opposition: This kind of opposition separates ingroups from outgroups by providing the edge and describing the features of both categories. This distinction could be based not only on opposing social groups—ethnic, religious, cultural—but also on ideology groups and social development levels.
- Mythic narratives: stories about the foundation and development of the nation. The structures of these myths could be reduced to several patterns: impediment by outgroups, condemning imposition, positive ingroup predisposition, validation of rights, enlightening, opposite interpretations of the same subject, and the same interpretation of opposite subjects.
- Normative order: judgments about the desirable state system and relations between the ingroups. Different binaries and myth structures are based on cultural rights, acceptance by advantaged and disadvantaged groups, legitimizing ingroups and delegitimizing outgroups, validation of social order, and consensus among groups.

This detailed structure is sharpened for the national narrative, but we can assume that other identity narratives (political, religious, gender, professional) have a similar structure.

In addition to identity narratives, there is also a more applied type of action narrative. We operate from the notion that there are pairs of interconnected narratives of identity and action. For example, the strategic narrative is a kind of action narrative that subordinates to the national narrative during a war (Jilani, 2020). Similarly, the electoral narrative (also action narrative) subordinates the political ideology narrative during a campaign (Zakharchenko and Zakharchenko, 2021). We can assume that every identity narrative can be paired with an action narrative during some special, challenging periods. These action narratives necessarily contain calls to action, for example, in the case of an election narrative, to vote for the candidate, campaign for him, persuade the opponents, and so on.

This idea of pairs of identity/action narratives allows us to elaborate on the approach for the structural analysis of the strategic narrative based on such patterns for the national narrative. We will describe it in Chapter 2.

2.2 Strategic narratives and their features

To answer RQ1 and RQ2, we first need to find out what approaches exist to studying strategic narratives. Its concept developed on the edge of centuries. In the context of the war, it is defined as a «form of communication, through which political actors attempt to give meaning to the past, present and future to achieve political objectives» (Szostek, 2017a, p. 5). National and strategic narratives are considered second-order narratives in the classification (Roselle et al., 2014), while the researchers placed the narratives of international organizations on the first level of the world's system of narratives.

There are several approaches to defining the goals of strategic narratives. J. Szostek distinguishes promotional measures of the strategic narrative, such as "nation-branding," aimed at attracting citizens and protecting the state's identity, and defensive measures directed against opponent states (Szostek, 2017a, p. 18). L. Freedman considers narratives that serve strategies aimed at hostile out-groups to persuade them to deceive or confuse and narratives about strategies aimed at in-groups and their mobilization (Freedman, 2015, p. 22). It is also believed that strategic narratives could be used not only for winning and domination but also for the establishment of cooperation (Miskimmon et al., 2016).

War in Afghanistan became the broadest polygon for strategic narrative studies (De Graaf, 2015). Scholars realize that communications with opponents, the local population, and home actors should be analyzed (Dimitriu, 2012). Nevertheless, researchers in this war have focused most on studying the strategic narrative within the United States, not in Afghanistan.

A different situation is the study of the Russian strategic narrative and its impact on the European Union and Ukraine. It was studied as offensive; for example, the narrative analysis of political papers showed the danger from Russia to Western democracies in 2017 (Miskimmon and O'Loughlin, 2017). The Ukrainian polygon brought to the fore the limitations of this "weapon" (Szostek, 2017b).

The Chinese strategic narrative has recently been considered an offensive weapon (Ichihara, 2020; Gustafsson and Hagström, 2021).

To answer RQ2, we should also examine the "recipes" of a good strategic narrative. First, it must meet the requirements of storytelling: to have a character or actors, setting and environment, conflict or action, and resolution (Roselle et al., 2014). It should be "a compelling storyline that can explain events convincingly and from which inferences can be drawn" (Freedman, 2006). Emotions and metaphors are very important, along with evidence or experience.

Despite the internal narrative features, external ones are also important. A good offensive strategic narrative has to resonate with local political myths. For example, it is argued that the Russian narrative was so influential in France because of its resonance with some French myths like the "Golden Age," "American danger," "European civilization," and so on (Schmitt, 2018). Another external requirement is that the narrative must not be disproven by further events and new information that could appear (Freedman, 2006, p. 23). Moreover, a narrative will be more convincing if it has an internal imperative and will be more salient if it catches attention (Freedman, 2015, p. 24).

At last, Dimitriu and De Graaf (2016) summarize the "strong narrative" features based on previous studies, particularly (Ringsmose and Børgesen, 2011). The narrative should be:

1. Clear, realistic, and with a compelling mission purpose.
2. Legitimate, in both objective (judicial, procedural) and subjective (political, public, ethical) senses.
3. Promising wartime success.
4. Presented consistently, and preferably without strong counternarratives.
5. Fitting within an overall strategic communication plan.

We can expand on the last point: this narrative must fit into an overall strategic plan that includes both information and conventional military operations.

As we can see, all these requirements tell more about the final features of strategic narratives and do not help to construct them. This is not surprising because scholars usually consider strategic narrative as an art rather than a construct: "The art of crafting strategic narratives is much more than a PR trick to "sell a war" or a mere tool for communication specialists," write De Graaf (2015).

This study will attempt to show what this structure might look like.

2.3 Affective strategic narrative

There is a strong belief that authorities can only develop strategic narratives. "Strategic narratives do not appear out of the blue, however. They are deliberately designed and nurtured by political elites," De Graaf says (De Graaf, 2015, p. 8; Freedman, 2006), strategic narratives are "deliberately constructed or reinforced out of ideas and thoughts that are already current." Only after the creation of the narrative do elites utilize different channels for its promotion, like celebrities (Wright, 2021), media, or opinion leaders. That is why most strategic narrative studies analyze official documents, leaders' speeches, or at least the rhetoric of propagandist media. Social media was studied as a source of counternarratives produced by informal leaders (Hellman and Wagnsson, 2015) or for less legitimate structures like ISIS (Siegel and Tucker, 2018).

The same belief was common among the scholars who studied protest activity: according to classical approaches, each large-scale and long-term protest should have a core organization. However, in the last decade, "social media protests" have undermined this belief, like the Revolution of Dignity in Ukraine or the "Arab Spring" in the Middle East. In this type of protest, source mobilization occurs due to the "weak ties" established by social media instead of the "strong ties" inside the organization (Metzger and Tucker, 2017). Under such circumstances, a completely new type of communication takes place. Z. Papacharissi calls it the "affective public" (Papacharissi, 2015). People who participate in affective social media movements provide their own "connective gatekeeping" of the news, as well as "connective framing" and even "connective storytelling," in which the same people can be its narrators and its heroes as they participate in the events that describe them in social media.

A very similar situation occurred during the first months of the full-scale war in Ukraine on 24 February 2022. As Zakharchenko (2022b) proved, Ukrainian social media users' behavior during the recent war was similar to the connective public. These users created connective messages and stories about the war and elaborated on the goals of the war. It is shown that neither Ukrainian nor Russian authorities had a strong impact on the social media content about the war, and often, the mode of this content was completely different from the official messages. Therefore, the social media narrative about the ongoing Russian–Ukrainian war is created in the mode of affective public.

RQ1 was derived from the notion that this affective war narrative has features of the strategic narrative because the online community that produces it aims to win the war. This narrative was successfully used for this purpose. This phenomenon can be named "connective strategic narrative."

This type of strategic narrative challenges its researchers (Roselle et al., 2014). proposed a three-step research of strategic narratives, including a study of their formation, projection, and reception. However, in the case of a connective strategic narrative, these three stages converge as their creators, distributors, and recipients are generally the same people.

It is important to know that the Russian strategic narrative proliferates oppositely and more commonly. Its dissemination is highly centralized, which is typical for authoritarian regimes. There are two core parts of the Russian propaganda machine: the first is Russian television, especially Channel One (Khaldarova and Pantti, 2016), and the second is the "troll factory"—an organization called the Internet Research Agency (Tucker et al., 2018) that spreads social media content on sensitive topics in different countries (Broniatowski et al., 2018; Committee on Foreign Relations United States Senate, 2018), particularly Ukraine (Golovchenko et al., 2018).

2.4 Ukrainian social media environment and the war

In studying Ukrainian strategic narratives, it is also important to understand the environment in which it emerges.

According to the classification of Hallin and Mancini (2004), Ukraine has a polarized pluralist model of the media system that is typical for hybrid political regimes. Before the full-scale Russian invasion in 2022, TV channels were controlled by oligarchs who competed with each other, creating a comparatively free media environment. Besides the oligarch control, there were a lot of issues that hampered the development of the completely democratic model (Orlova, 2016, p. 457), including a small advertising market, unfinished public broadcasting reform, the problem of personal security of journalists in a time of the Ukrainian–Russian war, the hidden advertising practices, and, at last, Russian propaganda pressure (Peisakhin and Rozenas, 2018). This pressure was the reason for banning three pro-Russian TV channels in 2021.

After the full-scale war with Russia began, Ukrainian authorities mobilized information sources for the struggle: six of the most popular TV channels launched the so-called «United News» marathon, which provided unified coverage of the war. The audience of some oppositional channels was artificially limited.

Under such circumstances, social media is a very important environment for information exchange. Ukraine became a known polygon for social media and people's activity studies due to its deep traditions of activism. Due to this, Papacharissi's affective public has been formed on Ukrainian social media over the past decade. The first case was the Revolution of Dignity 2013–2014, which appeared due to online people activity (MacDuffee and Tucker, 2017). Then, during the subsequent Ukrainian-Russian war in Donbas and Crimea (the so-called limited war that started in 2014 and finished in 2022 with the full-scale invasion), a powerful online movement of volunteers and activists appeared (Ronzhyn, 2016). Particularly, it was a counter-propaganda movement that was carried out not by government agencies but by commercial and activist organizations (Bolin et al., 2016). Even ordinary citizens joined the information resistance by creating numerous memes (Makhortykh and Sydorova, 2017). The experience of the Ukrainian women's movement in social media was also unique: the hashtag about the violence against women #янебоюсьсказати (#iamnotafraidtosayit) took place in Ukraine a year before the #metoo movement (Lokot, 2018). At last, the 2019 presidential election also had the format of affective discussion (Zakharchenko et al., 2019). Hostility between the supporters of former president Poroshenko and incumbent Zelensky, on the one hand, weakened the whole patriotic movement, but on the other, engaged new people in this movement that became helpful during the ongoing war (Zakharchenko and Zakharchenko, 2021). This is the communication landscape for the deployment of the Ukrainian–Russian information war.

3 Approach and methods

3.1 Data source

As a semi-processed source of the content for the detection of the narratives, we used two lists of key messages, which were used by the patriotic Ukrainian accounts and pro-Russian accounts in the war communication on social media.

These lists were created by the volunteering group CAT-UA (Communication Analysis Team - Ukraine, 2024). This dataset is not publicly available, as the group mainly provides services to the Ukrainian government, but we have obtained access to it. Since 24 February 2022, CAT-UA analyzed social media content about the war in Ukraine and provided the results for Ukraine's military and civil authorities. Every day, this group manually coded posts collected by one of the commercial media monitoring systems.

These are posts from Facebook, Instagram, Twitter, YouTube, Telegram, Vkontakte, and TikTok, found on a query containing Ukrainian and Russian words describing military actions. CAT-UA took into account posts (not comments or reposts) with the geolocation "Ukraine" or "Not defined." There were, on average, 320,000 such posts daily. A total of 1,050 daily messages were randomly selected for coding, or 42,000 in total, during the 24 February—4 April.

This organization used several coding categories, but for the purposes of this study, important three of them: attitude to the war (patriotic Ukrainian or pro-Russian attitude), link to the official source, and key message of the post. The first category was determined by coders based not only on the content of the post but also on the content of the author's account. They took into consideration explicitly expressed attitudes to the war in the profile or in the recent posts, the visual design of the profile, lexical features, and so on. Direct or indirect links to statements of the central or local, civil or military authorities of Ukraine were the criteria for the second category in the case of patriotic authors, or, in the case of pro-Russian accounts, to Russian officials, "officials" of occupation administrations, and official Russian propagandist media. Regarding the third category, CAT-UA coders used the methodology of PR message analysis (Zakharchenko, 2022a). This method was developed for PR campaign analysis but was successfully used for war communication and propaganda analysis. Therefore, the coder determines a logical statement that is substantial for communication and in which the subject or the predicate is related to the war. To avoid confusion with social media posts, which are also often referred to as "messages," we will refer to such judgments in this article as "communication messages."

Formulations of such communication messages are unified to emphasize the general meaning of each post. Examples of messages used by patriotic Ukrainian users are: «Ukrainian army repulses Russian attacks. Russians have heavy loses», «Ukrainian cities are sheltering», «Life is going on despite the war», «Ukraine captures or destroys Russian military equipment,», «There is an opportunity to support financially Ukrainian army», «Russian military loses grow», «Ukrainians help each other during the war». And here are examples of messages used by pro-Russian users: «Ukraine is shelling civilians in the Luhansk and Donetsk People's Republic», «Armies of Luhansk and Donetsk People's Republic advancing on Ukrainians», «Nazis & extremists operate in Ukraine», «Ukraine spreads fakes», «Russia takes control of cities and strategic objects», «Western sanctions will threaten world economics». As none of the posts was linked to its author, and all the aggregated data is in the anonymized form, there is no ethical violation.

This list of communication messages, without regard to their quantitative distribution, became the basis for our qualitative analysis of pro-Ukrainian and pro-Russian narratives about the war.

In total, 681 pro-Ukrainian communication messages were detected by CAT-UA coders in the dataset, and 172 were pro-Russian communication messages. Among them, 389 pro-Ukrainian communication messages were used only in unofficial communication, so these statements were made only by bloggers, journalists, independent experts, or, mostly, by ordinary social media users. In the pro-Russian dataset, there were only 16 completely unofficial communication messages.

3.2 Structural analysis of the strategic narrative

We used the dataset described above as the basis for our own coding based on the matrix of the strategic narrative structure, which we will develop based on the following.

Analytical framing of our research questions requires a shift from the classical approach to the strategic narrative as a masterpiece and, instead, an elaboration of a clear model for structural analysis of the strategic narrative.

As demonstrated in section 1.1, the national narrative is closely linked to the strategic narrative, forming a pair of identity and action narratives. Therefore, we can use Korostelina's structural model to develop one. When tailoring this model, we must consider the purposive nature of the strategic narrative.

Under the conditions of war, binary opposition turns into disposition, which describes not just ingroup and outgroup but at least four categories: "we," "enemies," "allies," and "others." We will refer to this part of the narrative as "disposition."

As an explanatory part of the narrative, the normative order also undergoes modifications in the strategic narrative. It defines the goals of participation in the war for the military, civilians, and the whole country, as well as for the enemy, allies, and others.

Mythic narratives are also important as they describe how the war is conducted, combining individual facts into stories about the war that differ depending on which side is telling them.

A fourth part of the strategic narrative, which is not present in the national narrative, is a call to action.

These four components of the strategic narrative, which are further subdivided into subcomponents, form the basis of our strategic narrative analysis matrix, shown in Figure 1.

Based on this scheme, the author provided topical coding of the communication messages in two lists used by pro-Russian and patriotic Ukrainian users on social media, assigning them to one of the hierarchical categories shown in Figure 1. Consequently, communication messages used in social media were assigned to the matrix of the strategic narrative.

To accomplish this, sets of communication messages representing each category were obtained. These sets were then qualitatively processed and reformulated to obtain a descriptive narrative structure for each category. At this stage, completed narratives used by both sides of the war were stated without considering oppositional and other side narratives. For instance, some users in Crimea or Donetsk shared a small number of posts from Russian users convicting Putin of the war. However, these communication messages were not part of the completed pro-Russian narrative used by Russians to bring victory closer, nor were they part of the Ukrainian narrative because they were not created by the Ukrainian side for defense. Therefore, such posts were not considered.

4 Results and discussion

The result of our coding and further transformation of the received data into the narrative form is presented in the qualitative dataset associated with this article.[1]

1 https://figshare.com/s/435ce9089fdf9cf219a3

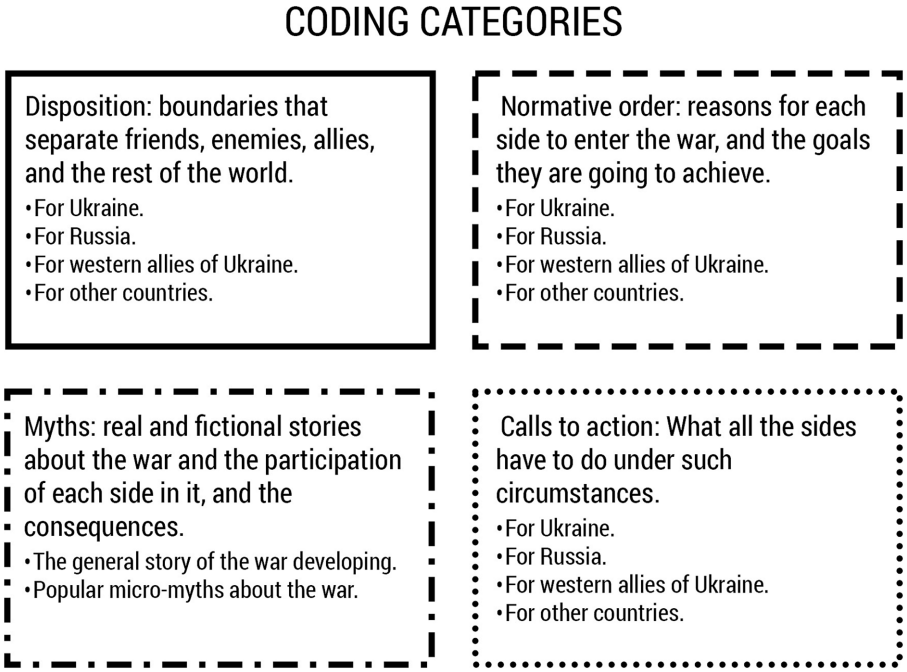

FIGURE 1
Matrix of the strategic narrative structural analysis.

Below, we present a short, illustrated description of these narratives along with their comparison, and then, we will provide a more detailed discussion about the aim of the study.

4.1 Comparison of patriotic Ukrainian and pro-Russian social media narratives in Ukraine

4.1.1 Disposition

A brief description of the dispositions presented in the two social media narratives studied is shown in Figure 2 (Ukrainian narrative) and Figure 3 (Russian narrative). These texts are abbreviated versions of the dispositions presented in the qualitative dataset mentioned at the beginning of section 4.

From this, we can see that both Ukrainians and Russians depict the disposition in a similar way: their own army is shown as powerful and supported by people, and the enemy as weak. Both describe the Ukrainian army as a unity of people rather than a leadership structure, unlike the Russian army, which looks like Putin's instrument. Western countries in both narratives are depicted controversially but with different accents. The Ukrainian narrative focuses more on other countries, while the Russian one does not pay enough attention to it (Figure 3).

4.1.2 Normative order

A brief summary of the regulatory procedure is also presented in Figure 4 (Ukrainian narrative) and Figure 5 (Russian narrative).

Both sides of the narrative (see Figures 4, 5) speak about their own goals on a global scale (to protect the world against totalitarianism vs. to destroy NATO), along with ascribing to the enemy more local goals. Ukrainians suspect that Russia aims to eliminate all Ukrainians, whereas Russia provides contradictory versions of Ukrainian and Western goals. Ukrainians are depicted as both active and passive actors in this story.

4.1.3 Mythic part

In this sub-section, we will examine the characters of large and small war stories, as well as the general plot structure of these stories. The plots are presented in more detail in the datasheet.

The sides of the war focus on different facts; namely, Ukrainians speak more about the grassroots movement and heroism of ordinary people, while Russians speak more about the people of Donbas and their suffering in the past years. Additionally, at the fact level, Russian and Ukrainian stories of war have different structures. Similar to Korostelina's typology of mythical parts of national narratives (see section 1.2.), it is possible to compare strategic narratives with some classical plots. After the beginning of April 2022, the Ukrainian strategic narrative took the form of the "Middle-earth war" of absolute good and absolute evil, as told by Tolkien. Ukrainians even widely use the word "orcs" to name Russian soldiers. The Russians themselves use the aesthetics of the "Great Patriotic War," but the shape of their narrative is not similar to those used by the Soviet Union in the II World War because now Moscow is an aggressor, not a victim. Therefore, its strategic narrative is the missionary "Crusade," with the aim of liberating the "holy city" of Kyiv from "infidels."

4.1.4 Calls to actions

Finally, calls to action from both narratives are summarized in Figure 6 (Ukrainian narrative) and Figure 7 (Russian narrative).

As we can see, the Ukrainian narrative includes clear calls to three categories of people, the most pronounced for Ukrainians, in contrast to the pro-Russian narrative with only one call for Ukrainians and Western countries—to look down—and a poor set of calls to Russians.

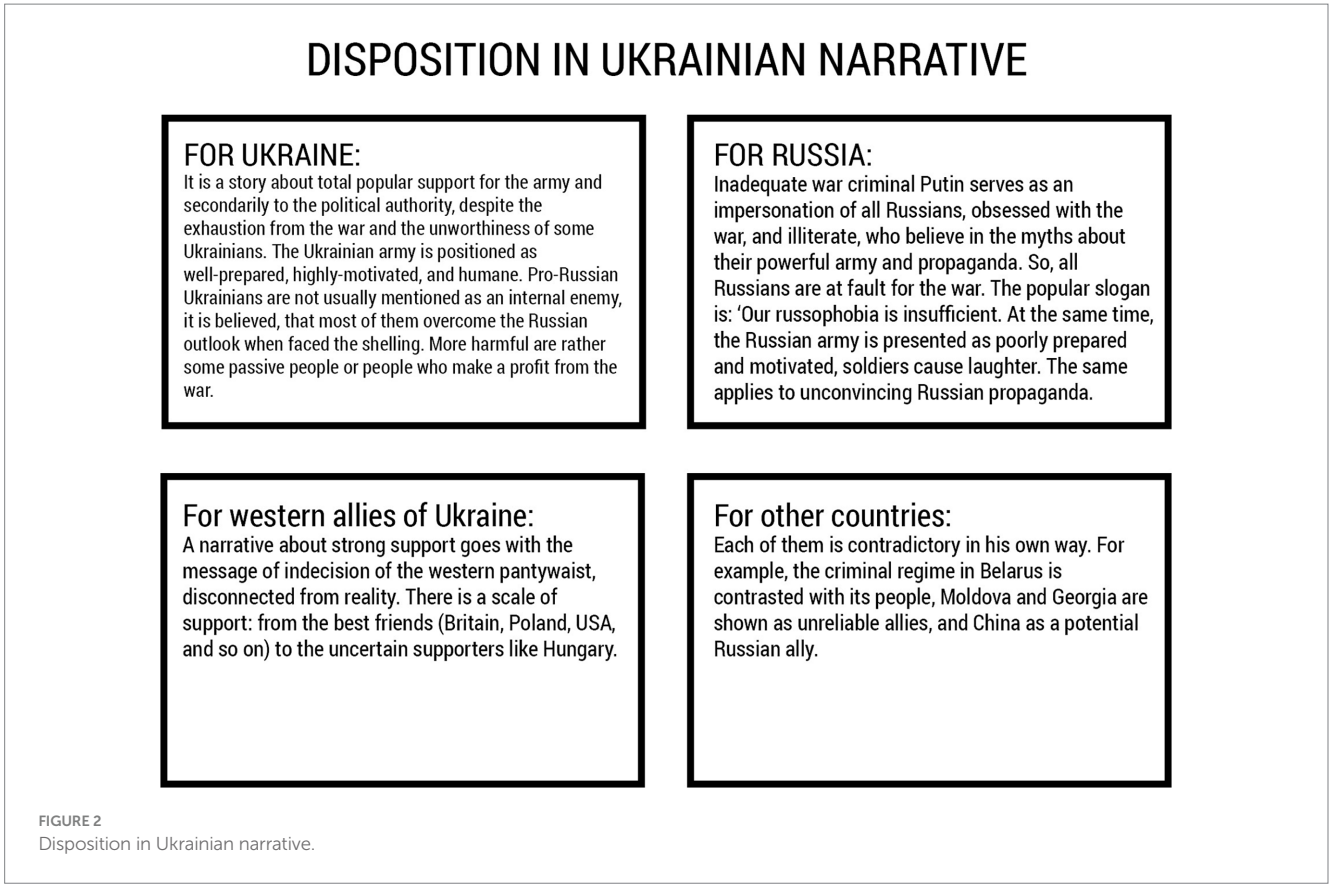

FIGURE 2
Disposition in Ukrainian narrative.

FIGURE 3
Disposition in Russian narrative.

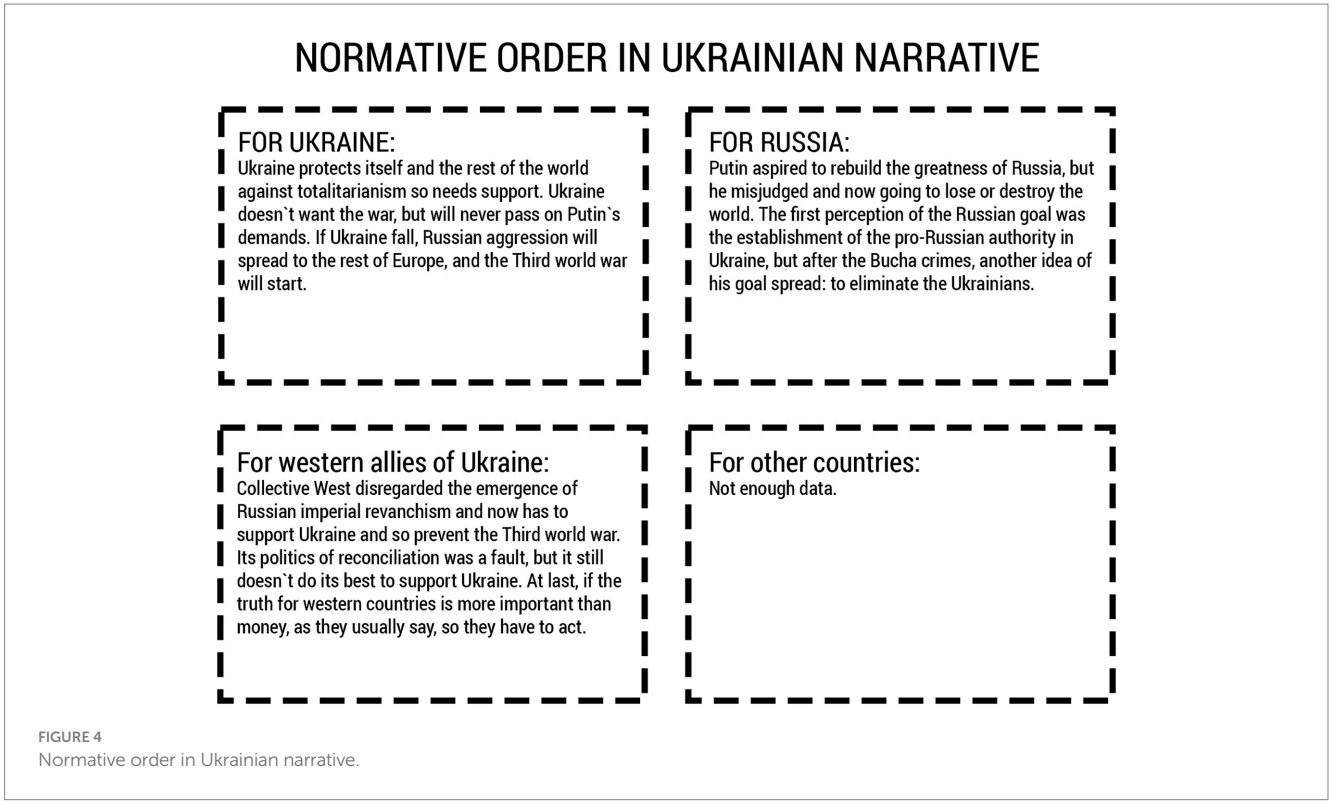

FIGURE 4
Normative order in Ukrainian narrative.

FIGURE 5
Normative order in Russian narrative.

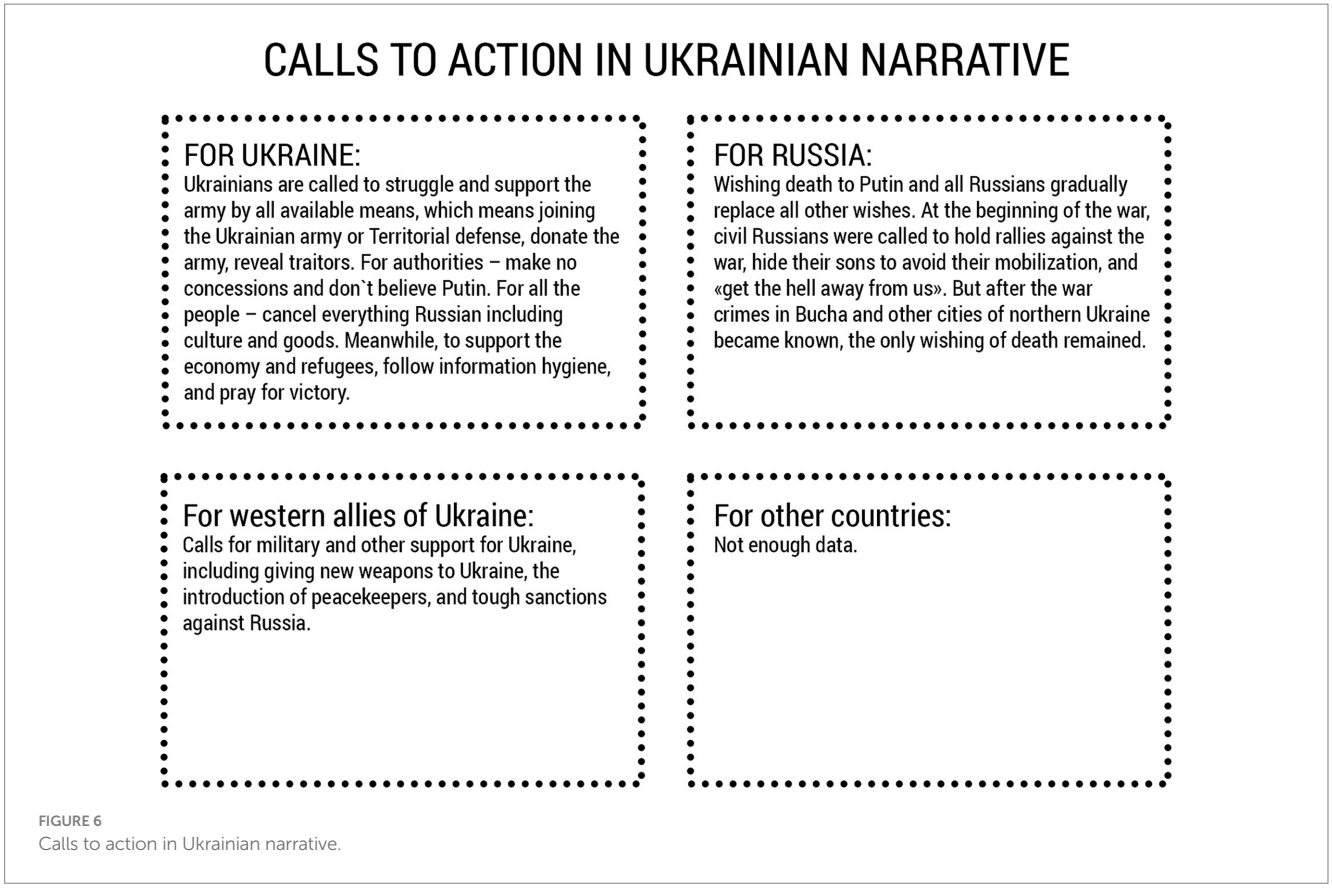

FIGURE 6
Calls to action in Ukrainian narrative.

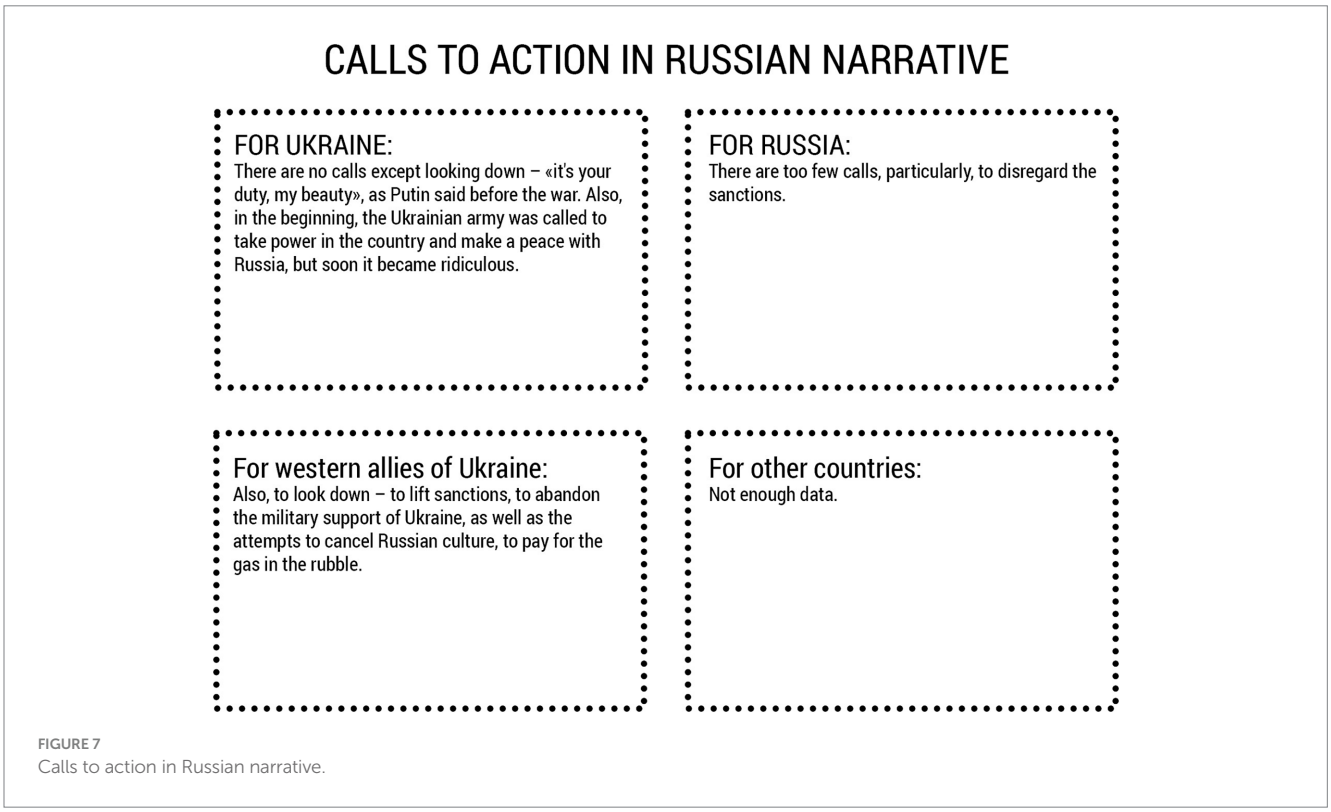

FIGURE 7
Calls to action in Russian narrative.

4.2 Social media narrative as a strategic narrative

As we can see from the previous subsection, the Ukrainian narrative in social media fills almost all cells of our strategic narrative matrix. Additionally, as noted in section 2.1, around 57% of the messages used in the social media narrative were not expressed by official sources. Let us examine how important these messages were.

Grassroots communication significantly influenced three of the four aforementioned sections of the war narrative. Only the normative order section was primarily composed of official communication.

Unofficial communication had the most significant impact on the mythical part of the narrative. Firstly, folklore stories about voluntary resistance to the occupiers make a significant contribution. True stories include the one about the tractor pulling the Russian tank, fictional stories about the old woman who brought down the drone with a jar of pickles, or another old woman who poisoned the occupants with pies. Secondly, ordinary people enriched this narrative with stories about life, self-organization, and resistance under occupation, siege, or shelling. For example, numerous stories about performances or concerts in bomb shelters, about volunteers who helped the besieged Chernihiv, and about peoples` protests in the occupied parts of Kherson and Zaporizhzhia regions.

Another section of the narrative heavily influenced by grassroots communication is a disposition, primarily focusing on judgments about Ukrainians and Russians. Messages about Western countries were under the control of authorities. Ordinary people began to talk about the changes in Ukrainians since the beginning of the war, notably the uprise of proactivity, mutual support, and self-awareness. This is where the image of the internal enemy emerged, including people who make money from the war, and, on the other hand, excluded the Russian-speaking population and even formerly pro-Russian people who are thought to have changed their views after being shelled. Additionally, within the unofficial segment, the statement that Russia is a terrorist state emerged, which was then repeated by the authorities in international communication, but after the period of this research. The image of Russia as an ever-present historical danger was also articulated. Finally, the image of "good Russians" who speak about their suffering from the war instead of helping Ukrainians emerged in the unofficial segment.

The latter and probably the most important narrative segment shaped by unofficial communication, is the call-to-action part. Mostly, the calls to help the army with buying military equipment appeared here because the authorities were not inclined to acknowledge the problems with ammunition. The same goes for the calls to boycott the companies that still work in Russia and to avoid internal struggles for some time. The calls to authorities to make no concessions and not to believe Putin were essential, and after the revelation of the war crimes in Bucha, even calls to withdraw from negotiations with Russia. At the same time, in the connective narrative, the official calls to Russian people to protest against the war or to avoid mobilization disappeared. These calls were substituted by emotional pleas for Russians to die, which determined further attitudes toward the negotiations.

This description does not include messages that appeared in the grassroots segment but were then repeated by authorities within the 40-day study period.

The narrative created by the Ukrainian affective public in social media serves as a true defensive strategic narrative. This public involves not only authorities but also top bloggers, media, ordinary military personnel, and civilians in the process of narration. Therefore, authorities appear to be just one of the communicators, albeit a very powerful one. For example, this communicator is responsible in this system for messages about relations with Western countries and goal setting. Moreover, the use of this affective narrative allows the Ukrainian side of the war to include people who do not trust official authorities.

Official communication does not completely provide the narrative that motivates Ukrainians to struggle. Conversely, the social media narrative includes both "official" and "unofficial" parts of such a strategic narrative. This affective public does not only quote the "official" and "unofficial" thoughts about the war but also comprehends these thoughts, produces new messages in passing, and attaches new framing. Sometimes, authorities have to co-opt the messages from the unofficial part into their formal communication.

Lastly, let us recall J. Szostek's definition of strategic narrative as "a form of communication through which political actors attempt to give meaning to the past, present, and future to achieve political objectives" (Szostek, 2017a, p. 5). Considering the affective public and civil society as political actors, we should regard the social media narrative about the war during the research period as strategic. This is the answer to RQ1.

The situation differs from that of the pro-Russian social media narrative in the Ukrainian information space. While it also covers almost all cells of the strategic narrative matrix, only about 9% of the messages used in this narrative were not expressed by official sources. These mostly include complicated statements by cultural figures, such as "This is the war between Western totalitarianism and Eastern traditionalism," or local observations about war episodes told by pro-war bloggers, such as "Ukrainian armed forces are trying to evacuate the defenders of Mariupol" or "Russia diverts troops from separate locations." Therefore, even in the occupied territories, the pro-Russian population in Ukraine mostly repeats official messages and does not produce its own messages.

This is why the pro-Russian social media narrative is simply a mirror of the official Russian strategic narrative. Russia conducts offensive information operations in the foreign information space and uses social media as the main source of disseminating such narratives, as other media channels are closed to them. Hence, this narrative in social media is precisely the narrative that Russia employs for information offensives. Therefore, it can also be used for this study, albeit for reasons different from the Ukrainian narrative. This is the answer to RQ2.

Positive answers to both RQ1 and RQ2 give us the right to consider section 4.1. as a response to RQ3.

4.3 Compliance with the requirements for narratives

Let us discuss the obvious problems for the Russian offensive narrative regarding the requirements listed in subsection 1.1, which will be the answer to RQ4.

4.3.1 Set of characters

The Russian set of characters is poor; it includes a collective enemy—"nationalist battalions"—with no prominent protagonists except Putin and his spokespeople. The Russian army was faceless in the first stage of the war, especially in comparison to the Ukrainian narrative, which has a rich set of protagonists that includes not only President Zelenskyi but also a lot of heroes from the people: old women or half-criminal actors who combat occupants, legendary combatants like the "ghost of Kyiv," and so on.

4.3.2 Presence of internal imperative

Presence of internal imperative: The pro-Russian narrative has no clear call for Ukrainian citizens and Western countries besides giving up and allowing Russians to do everything they want. The Ukrainian narrative has clear calls (see Figures 6, 7).

4.3.3 Readiness to further events and new information

The first version of the pro-Russian narrative, which included the concept of a "brother nation" that waits for Russian liberators, was disproved by reality. The Ukrainian narrative also underwent substantial changes caused by new events like the information about the Bucha tragedy, but these events did not change it crucially; they just made it more radical.

4.3.4 Presented consistently

The pro-Russian narrative has fundamental inconsistencies, for example, uncertainty about who was the initiator of the war—Ukraine or the West.

4.3.5 Inclusion of the narrative

The pro-Russian narrative is exclusive not only to patriotic Ukrainians but also to the 'fifth column' inside Russia that had been tempted by the Western lifestyle as well. The Ukrainian narrative has different "friend-foe" criteria: Ukraine does not condemn its citizens who have or had a pro-Russian outlook as long as they did not commit a crime like fire adjustment. More convicted is a category of unconscious people who make money from the war. Moreover, the Ukrainian narrative includes the internal political opposition, including supporters of former President Poroshenko, who do not support Zelenskyi but strongly support the Ukrainian army. This confrontation undermined the unity of society during the war but, conversely, makes military goals independent of the person of a political leader: potential delegitimization or even elimination of the president will not stop the resistance. In the pro-Russian case, much more is tied personally to Putin.

As a result, we can see that at the beginning of the war, when the disposition was set for further use, as well as the goals of the war, Russia did not have a solid strategic narrative.

4.4 Possible reasons for the low quality of the Russian narrative compared to the Ukrainian one

All of the problems with the Russian offensive narrative do not necessarily mean that nobody will believe in it. For example, some people may have strong emotional bonds with the traditional Russian national narrative and thus be receptive to strategic narratives created on its basis. However, it makes engaging new recipients and persuading them almost impossible. If Ukrainians had not believed in the Russian neo-imperial messages before, especially if they had contact with a coherent and compelling Ukrainian narrative, they would have been unlikely to become Russian supporters.

The reasons for this situation appear similar to those in conventional war: miscalculations based on problems with feedback within the Russian elites. Jeremy Fleming, the head of the British agency GCHQ, said that Putin misjudged the strength of Ukrainian resistance, the Western response, and the ability of his forces to deliver a rapid victory because his advisers were afraid to tell him the truth.

Another reason is the differences in the strategic narrative creation process. The affective public can be more adaptive to reality because it has rapid feedback from actual events. For example, at the beginning of the full-scale invasion, good news from the battleground became an important part of the Ukrainian strategic narrative as proof of Kyiv's ability to confront Russia. In the case of a highly centralized Russian bureaucratic structure, information flows, including feedback, are much slower than in a decentralized one, and there are other artificial obstacles in the way.

There may also be another reason for the inconsistency of the Russian narrative, as highlighted by T. Snyder (Boborykin, 2022): the Russian government believes in the postmodern relativity of facts and uses this approach for its propaganda for different audiences. Snyder believes that, in this case, a meaningful story told by Ukrainians is much more powerful than "literary criticism" of Russian propaganda.

4.5 Challenges for the connective strategic narrative

The situation when state structures have lost their monopoly on the strategic narrative has both advantages and disadvantages for the country's information sustainability. The advantages were illustrated in the previous sections of this article: firstly, public creativity allows it to be more diverse and adaptive to the challenges of reality, and secondly, its grassroots nature allows it to better respond to the moods of citizens and their expectations.

As for the challenges, they are inherent in the very nature of connectivity.

- Sensitivity to democracy. The grassroots activity of citizens in creating content about the war is stimulated by the feeling that citizens can influence the approach to victory. This feeling is directly opposite to the paternalism that prevails in totalitarian and post-totalitarian societies. Therefore, in such circumstances, the government cannot contradict the most widespread public beliefs but rather needs to successfully complement and guide them. This is hindered by the challenges faced by democracy during the war, including military censorship, restriction of the rights to protest, to travel abroad for men, and so on. In such circumstances, it is important to introduce only those restrictions that the majority of citizens are willing to accept and to explain in detail the need for such restrictions. Otherwise, distrust of the authorities naturally begins to form, and hence, the feeling that the goals of citizens are not in line with those of the authorities. In such circumstances, there is a discrepancy between the strategic narrative created by the authorities and the one shaped by society.
- War fatigue is also naturally linked to trust in the authorities. Given that the state of affect is exhausting and leads to natural fatigue, its duration is limited in time. All the cases of social movements described by Z. Papacharissi, during which the affective public existed, lasted for several months. Therefore, when a war lasts long enough, there inevitably comes a time when the affective public ceases to exist. At this point, the

authorities should be prepared for the fact that a powerful generator of new ideas that fuels the strategic narrative will disappear in society. This means that they will have to continue shaping this narrative on their own. In addition, Papacharissi emphasizes that it is impossible to create a state of affective public artificially, so we cannot hope to restore the unity of the people that was observed at the beginning of the full-scale invasion.

- One of the signs of societal fatigue can be the manifestation of dividing lines that citizens may ignore for a while. At a certain point, however, unsolved social contradictions become impossible to ignore, and citizens begin to accuse each other of misbehavior during the war. This threatens the coherence of the strategic narrative, and "conflict copies" (NGO "CAT-UA", 2023) appear in it.
- Against this background, totalitarian systems with a centralized information environment, if well organized, can be more stable as they suppress the activities of critics of the authorities and unify messages between different speakers. However, this can only happen if all the challenges of totalitarian systems described in the previous section of this article are overcome.
- Propaganda can use modern technological solutions to influence information. In addition to creating and spreading fake news, for which artificial intelligence systems are often used (Huntsman et al., 2024), it is also possible to use AI to formulate and promote a strategic narrative. Such systems can formulate fictional stories about war heroes or help to present real stories in more attractive and diverse forms. They can also help formulate the goals of the war, prepare, and "package" appeals to different segments of society. One of the potential tasks for AI could be to reconcile the stories that already exist in the environment of influence and those that propaganda wants to spread. If used correctly, AI can challenge the affective public in creativity and thus in its potential to influence the motivation of citizens.

5 Conclusion

We have presented a novel approach to analyzing modern strategic narratives. Our findings suggest that strategic narratives can emerge not only from intentional construction by elites but also from the "affective public" on social media—emotionally tied users. This phenomenon can be named "connective strategic narrative." We have also developed a structural pattern for strategic narrative analysis that enables comparison and identification of strengths and weaknesses.

We proved that the Ukrainian defensive strategic narrative at the beginning of the full-scale invasion was formed in this way, unlike the Russian offensive narrative, which was formed centrally by the authorities and only spread on social media.

The patriotic Ukrainian narrative developed by the affective public is strategic because it is more detailed than the official Ukrainian narrative from authorities and more inclusive, incorporating both official information and additional mythic elements, disposition messages, and calls to action.

Our analysis reveals significant weaknesses in the Russian strategic narrative, including incoherence, a lack of compelling characters and imperatives, a lack of inclusiveness, and disconfirmation by new events. These shortcomings may explain the narrative's inefficiency during the war.

We have also shown that the "connective strategic narrative" has both advantages and disadvantages, including its short duration and, thus, additional challenges for the authorities or other social institutions that have to replace the affective public in time to update such a narrative.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

AZ: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the VolkswagenStiftung folgende under grant 9B 984.

Acknowledgments

I would like to offer my special thanks to the CAT-UA volunteering group that analyzes war communication in Ukraine and provides a dataset for this study. I would also like to thank the Armed Forces of Ukraine for providing security to fulfill this work. This article has become possible only because of the resilience and courage of the Ukrainian Army.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpos.2025.1434240/full#supplementary-material

References

Boborykin, A. (2022). Timothy Snyder: Russia calls itself a democracy, but it's obviously not | Ukrainska Pravda: Ukrainska Pravda September 15. Available at: https://www.pravda.com.ua/eng/articles/2022/09/15/7367504/ (Accessed January 3, 2025).

Bolin, G., Jordan, P., and Ståhlberg, P. (2016). "From nation branding to information warfare: Management of Information in the Ukraine-Russia conflict" in Media and the Ukraine crisis: hybrid media practices and Narratives of conflict. ed. M. Pantti (New York, Bern, Berlin, Bruxelles, Frankfurt am Main, Oxford, Wien: Peter Lang), 3–18.

Broniatowski, D. A., Jamison, A. M., Qi, S. H., AlKulaib, L., Chen, T., Benton, A., et al. (2018). Weaponized health communication: twitter bots and Russian trolls amplify the vaccine debate. *Am. J. Public Health* 108, 1378–1384. doi: 10.2105/AJPH.2018.304567

Burns, R. (2022). Russia's failure to take down Kyiv was a defeat for the ages. *AP News*. April 7. Available at: https://apnews.com/article/russia-ukraine-war-battle-for-kyiv-dc559574ce9f6683668fa221af2d5340 (Accessed January 3, 2025).

Caliskan, M. (2019). Hybrid warfare through the Lens of strategic theory. *Def. Secur. Anal.* 35, 40–58. doi: 10.1080/14751798.2019.1565364

Carmen, L. C., Marled, M. S. A., Gómez, D., and Luis, R. (2023). Social media narrative and its prosumers: a systematic review from 2016-2023. *Migr. Lett.* 20, 1004–1015. doi: 10.59670/ml.v20iS6.4542

Committee on Foreign Relations United States Senate (2018). Putin's asymmetric assault on democracy in Russia and Europe: Implications for U.S. National Security. Washington, DC: U.S. Government Publishing Office.

Communication Analysis Team - Ukraine. (2024). Available at: https://cat-ua.org/en/homepage/ (Accessed December 31).

De Graaf, B. (2015) in Strategic narratives, public opinion and war: winning domestic support for the afghan war. eds. G. Dimitriu and J. Ringsmose (London: Routledge).

Dimitriu, G. R. (2012). Public relations review winning the story war: strategic communication and the conflict in Afghanistan. *Public Relat. Rev.* 38, 195–207. doi: 10.1016/j.pubrev.2011.11.011

Dimitriu, G., and De Graaf, B. (2016). Fighting the war at home: strategic Narratives, elite responsiveness, and the Dutch Mission in Afghanistan, 2006-2010. *Foreign Policy Anal.* 12, 2–23. doi: 10.1111/fpa.12070

Freedman, L. (2006). The transformation of strategic affairs. London: Routledge.

Freedman, L. (2015). "The possibilities and limits of strategic Narratives" in Public opinion and war: winning domestic support for the afghan war. ed. S. Narratives (London: Routledge), 17–36.

Golovchenko, Y., Hartmann, M., and Adler-Nissen, R. (2018). State, media and civil Society in the Information Warfare over Ukraine: citizen curators of digital Disinformation. *Int. Aff.* 94, 975–994. doi: 10.1093/ia/iiy148

Gustafsson, K., and Hagström, L. (2021). The limitations of strategic Narratives: the Sino-American struggle over the meaning of COVID-19. *Contemp. Secur. Policy* 42, 415–449. doi: 10.1080/13523260.2021.1984725

Hall, B. (2022). Military briefing: Ukraine's battlefield agility pays off | financial times. *The Financial Times*. Available at: https://www.ft.com/content/9618df65-3551-4d52-ad79-494db908d53b (Accessed January 3, 2025).

Hallin, D. C., and Mancini, P. (2004). Comparing media systems. Cambridge: Cambridge University Press.

Hellman, M., and Wagnsson, C. (2015). New media and the war in Afghanistan: the significance of blogging for the Swedish strategic narrative. *New Media Soc.* 17, 6–23. doi: 10.1177/1461444813504268

Horbyk, R., Dutsyk, D., and Shalaiskyi, S. (2023). Effectiveness of Russian disinformation counteraction in Ukraine in a full-scale war. Kyiv: NGO Ukrainian Media and Communication Institute. Available at: https://www.jta.com.ua/wp-content/uploads/2023/08/UMCI_-Effectiveness-of-Russian-Disinformation-Counteration_EN.pdf (Accessed January 3, 2025).

Huntsman, S., Robinson, M., and Huntsman, L. (2024). Prospects for inconsistency detection using large language models and sheaves. ArXiv.

Ichihara, M. (2020). Is Japan immune from China's media influence operations? *The Diplomat*. Available at: https://thediplomat.com/2020/12/is-japan-immune-from-chinas-media-influence-operations/ (Accessed January 3, 2025).

Jilani, S. G. (2020). National narrative building: role of academia, think tanks and media. Islamabad: ISSRA Papers XII.

Katerynchuk, P. (2017). Russian propaganda as an instrument of foreign policy strategy towards Ukraine. *Історико-Політичні Проблеми Сучасного Світу* 33–34, 222–229. doi: 10.31861/mhpi2016.33-34.222-229

Kaun, A., and Fast, K. (2013). Research report: Mediatization of culture and everyday life: Stockholm Available at: https://www.diva-portal.org/smash/get/diva2:698718/FULLTEXT02.pdf (Accessed January 3, 2025).

Kaye, H. J. (2016). It's time to cultivate a new grand narrative: History News Network Available at: https://historynewsnetwork.org/article/163248 (Accessed January 3, 2025).

Khaldarova, I., and Pantti, M. (2016). FAKE NEWS. The narrative Battle over the Ukrainian conflict. *Journal. Pract.* 10, 891–901. doi: 10.1080/17512786.2016.1163237

Korostelina, K. V. (2014). Constructing the Narratives of identity and power: self-imagination in a young Ukrainian nation. Lanham: Lexington Books.

Lokot, T. (2018). #IAmNotAfraidToSayIt: stories of sexual violence as everyday political speech on Facebook. *Inform. Commun. Soc.* 21, 802–817. doi: 10.1080/1369118X.2018.1430161

Lyotard, J.-F. (1984). The postmodern condition: a report on knowledge. Manchester: Manchester University Press.

MacDuffee, M., and Tucker, J. (2017). Social media and EuroMaidan: a review essay. *Slav. Rev.* 76, 169–191.

Makhortykh, M., and Sydorova, M. (2017). Social media and visual framing of the conflict in eastern Ukraine. *Media War Confl.* 10, 359–381. doi: 10.1177/1750635217702539

Metzger, M. M. D., and Tucker, J. A. (2017). Social media and EuroMaidan: A review essay. *Slav. Rev.* 76, 169–191. doi: 10.1017/slr.2017.16

Miskimmon, A., Loughlin, B. O, Roselle, L., Miskimmon, A., Loughlin, B. O., Roselle, L., et al. (2016). Strategic narratives: a response. Crit. Stud. Secur. 3. Routledge: 341–344. doi: 10.1080/21624887.2015.1103023

Miskimmon, A., and O'Loughlin, B. (2017). Russia's Narratives of global order: great power legacies in a polycentric world. *Politi. Govern.* 5, 111–120. doi: 10.17645/pag.v5i3.1017

NGO "CAT-UA" (2023). Inciting conflicts on social media: how they evolve and intensify the divisions among us. Available at: https://cat-ua.org/en/2023/03/16/ukrainians-afraid-of-russian-fakes-less-and-less-3/ (Accessed January 3, 2025).

Orlova, D. (2016). Ukrainian media after the EuroMaidan: in search of Independence and professional identity. *Publizistik* 61, 441–461. doi: 10.1007/s11616-016-0282-8

Papacharissi, Z. (2015). Affective publics: Sentiment, technology, and politics. New York: Oxford University Press.

Peisakhin, L., and Rozenas, A. (2018). Electoral effects of biased media: Russian Television in Ukraine. *Am. J. Polit. Sci.* 62, 535–550. doi: 10.1111/ajps.12355

Pihlaja, S. (2011). 'When Noah built the ark…' metaphor and biblical stories in Facebook preaching. *Metap. Soc. World* 7, 87–102. doi: 10.1075/msw.7.1.06pih

Ringsmose, J., and Børgesen, B. K. (2011). Shaping public attitudes towards the deployment of military power: NATO, Afghanistan and the use of strategic Narratives. *Eur. Secur.* 20, 505–528. doi: 10.1080/09662839.2011.617368

Ronzhyn, A. (2016). "Social media activism in post-Euromaidan Ukrainian politics and civil society" in Proceedings of the 6th International Conference for E-democracy and Open Government, CeDEM 2016. Krems, Austria: Danube University, 96–101.

Roselle, L., Miskimmon, A., and Loughlin, B. O. (2014). Media, war and conflict strategic narrative: a new means to understand soft power. *Media War Confl.* 7, 70–84. doi: 10.1177/1750635213516696

Schmitt, O. (2018). When are strategic narratives effective? The shaping of political discourse through the interaction between political myths and strategic Narratives. *Contemp. Secur. Policy* 39, 487–511. doi: 10.1080/13523260.2018.1448925

Siegel, A. A., and Tucker, J. A. (2018). The Islamic State's information warfare. Measuring the success of ISIS's online strategy. *J. Lang. Polit.* 17, 258–280. doi: 10.1075/jlp.17005.sie

Snyder, T. (2004). The reconstruction of nations. New Haven: Yale University Press.

Szostek, J. (2017a). Defence and promotion of desired state identity in Russia's strategic narrative. *Geopolitics* 22, 571–593. doi: 10.1080/14650045.2016.1214910

Szostek, J. (2017b). The power and limits of Russia's strategic narrative in Ukraine: the role of linkage. *Perspect. Polit.* 15, 379–395. doi: 10.1017/S153759271700007X

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., et al. (2018). Social media, political polarization, and political disinformation: a review of the scientific literature. *SSRN Electron. J.* doi: 10.2139/ssrn.3144139

Wright, K. A. M. (2021). NATO' s strategic Narratives: Angelina Jolie and the Alliance' s celebrity and visual turn. *Rev. Int. Stud.* 47, 443–466. doi: 10.1017/S0260210521000188

Zakharchenko, A. (2022a). PR-message analysis as a new method for the quantitative and qualitative communication campaign study. *Inf. Med.* 93, 42–61. doi: 10.15388/Im.2022.93.60

Zakharchenko, A. (2022b). The clash of strategic Narratives in the Russo-Ukrainian war. *Forum for Ukrainian Studies*. September 18. Available at: https://ukrainian-studies.ca/2022/09/18/the-clash-of-strategic-narratives-in-the-russo-ukrainian-war/ (Accessed January 3, 2025).

Zakharchenko, A., Maksimtsova, Y., Iurchenko, V., Shevchenko, V., and Fedushko, S. (2019). Under the conditions of non-agenda ownership: social media users in the 2019 Ukrainian presidential elections campaign. *CEUR Workshop Proc.* 2392, 199–219.

Zakharchenko, A., and Zakharchenko, O. (2021). The influence of the 'Tomos narrative' as a part of the Ukrainian national and strategic narrative. *Corvinus J. Sociol. Soc. Policy* 12, 163–178. doi: 10.14267/CJSSP.2021.1.7

frontiers | Frontiers in Political Science

TYPE Brief Research Report
PUBLISHED 07 February 2025
DOI 10.3389/fpos.2025.1517726

OPEN ACCESS

EDITED BY
Paul Vines,
Two Six Technologies, United States

REVIEWED BY
Ralph Schroeder,
University of Oxford, United Kingdom
Hugh Lawson-Tancred,
Birkbeck University of London,
United Kingdom

*CORRESPONDENCE
Federico Pilati
✉ federico.pilati2@unibo.it
Tommaso Venturini
✉ tommaso.venturini@unige.ch

RECEIVED 26 October 2024
ACCEPTED 14 January 2025
PUBLISHED 07 February 2025

CITATION
Pilati F and Venturini T (2025) The use of artificial intelligence in counter-disinformation: a world wide (web) mapping.
Front. Polit. Sci. 7:1517726.
doi: 10.3389/fpos.2025.1517726

The use of artificial intelligence in counter-disinformation: a world wide (web) mapping

Federico Pilati[1,2,3]* and Tommaso Venturini[3]*

[1]Department of Political and Social Sciences, University of Bologna, Bologna, Italy, [2]Department of Sociology and Social Research, University of Milano-Bicocca, Milan, Italy, [3]Institut des sciences de la communication et des cultures numériques "Medialab", University of Geneva, Geneva, Switzerland

Disinformation has recently become a subject of widespread concerns across the globe. To combat this issue, various initiatives have emerged, aimed at identifying, tracking, and debunking disinformation. Artificial intelligence (AI) has been incorporated as a tool to counter disinformation, but its implementation has not always been successful and may even be counterproductive. Thus, there is a growing recognition of the need for benchmarking the various ongoing efforts to ensure greater efficacy and coordination in the use of AI and assure that this does not lead to forms of algorithmic censorship. Our goal is to provide a mapping of the projects that use AI to counter disinformation by means of their hyperlink network analysis to shed light on their aims, approaches, and challenges.

KEYWORDS

artificial intelligence, disinformation, counter-disinformation, web mapping, hyperlink network analysis

Introduction

The proliferation of digital media and social networks sites has enabled the rapid spread of problematic information (Vosoughi et al., 2018), and traditional approaches to media regulation and censorship seem no longer sufficient to address this challenge (Alemanno, 2018; Marsden et al., 2020). Governments, academia and civil society are haunted by the idea of finding ways to combat this problem, and Artificial intelligence (AI) is increasingly being seen as an appealing tool in this fight. AI has indeed the potential to automate the identification of false or misleading information, which can then be flagged or removed before it can spread widely (Bontridder and Poullet, 2021). Organizations such as the European Union[1] and the United Nations[2] have launched initiatives to support the development of AI-powered fact-checking tools, while private companies like Meta[3] and Google[4] have invested in AI to help identify and remove false content from their platforms.

Although these top-down initiatives play an important role in addressing disinformation, they cannot operate alone. Indeed, a growing recognition of bottom-up actions that empower journalists and civil society organizations to combat disinformation is also on the rise (Golovchenko et al., 2018). Fact-checking initiatives have emerged as a critical player in the fight

1 https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news

2 https://www.itu.int/hub/2022/05/ai-can-help-fight-disinformation/

3 https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/

4 https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations

against disinformation, providing a valuable service to citizens and journalists alike (Graves, 2016; Porter and Wood, 2021). Many of these initiatives rely on AI to quickly identify and analyze large volumes of information and, in many cases, also develop their own AI-powered tools to enhance their fact-checking capabilities. As examples, *Full Fact*,[5] a non-profit fact-checking organization, has developed a review system that uses AI to identify claims made in political speeches and news articles and verify their accuracy. Similarly, *NewsGuard*,[6] a US-based company, has launched a tool for training generative AI services to recognize all the significant top false narratives spreading online and to use its ratings of web sources as signals to help both the machines and users of AI models to identify trustworthy news and information.

While there is no single solution to the problem of disinformation, AI has the potential to play a role in mitigating its impact (Kertysova, 2018). However, it is important that these efforts are transparent, carefully designed and implemented step-by-step to ensure that they are effective, and do not inadvertently harm free speech or democratic processes with forms of algorithmic surveillance and censorship (Marsden and Meyer, 2019; Gorwa, 2019). Presently, there has been a surge of research focused on leveraging machine learning to identify and flag false information, predict the virality potential of fake news, and provide comprehensive fact-checking and verification services (Choraś et al., 2021). By utilizing natural language processing (NLP) and machine learning algorithms, AI can analyze the language, sentiment, and structure of social media posts and news articles to detect patterns and identify potentially misleading or false content. AI can be also used to track the spread of disinformation and identify its sources to prevent it from spreading further. AI algorithms can analyze text, images, and videos to identify patterns and anomalies that may indicate disinformation. Furthermore, rather than be used in isolation (a use that is fraught with risks) AI can assist fact-checkers and journalists in verifying information, identify sources, cross-check information, and provide additional context to speed up the fact-checking process and improve accuracy (Bontcheva et al., 2024).

Recent research has demonstrated that cutting-edge language models can provide trustworthiness ratings for a diverse array of news outlets, accompanied by contextual explanations that align closely with human expert judgments (Yang and Menczer, 2023). This suggests a potential uptick in the adoption of these tools by fact-checking organizations in their ongoing efforts (Graves, 2018).

The continued advancement of AI-driven technologies has paved the way for more sophisticated approaches to combat the spread of disinformation. State-of-the-art machine learning models are now capable of discerning increasingly nuanced patterns in content generation and dissemination, allowing for more efficient identification and containment of misleading or false information. These developments hold promise for a future where AI not only aids in flagging and verifying the authenticity of information, but also actively contrasts the propagation of falsehoods using alerts across a diverse set of social media and search engines (Bontcheva et al., 2024). These two different approaches address the problem of disinformation at distinct points in the information ecosystem: the first approach tackles it downstream by identifying and managing misleading content after it has spread. The second approach works upstream, aiming to prevent the proliferation of falsehoods in the first place by delivering proactive interventions.

As these initiatives that use AI continue to evolve, it is essential to integrate them into a broader framework and a common ground. In fact, despite challenges and setbacks[7], increasing efforts are invested in AI as an assistant to help identify problematic information and counter the spread of disinformation (Graves, 2018). Taking roots from these promises, the objective of our research is to map the landscape of initiatives that use AI to combat disinformation.

Specifically, in this paper we analyze the hyperlink citation structure of the websites of the initiatives that use AI to fight disinformation. Leveraging on a mix of computational techniques and qualitative insights, we aim to identify and categorize these initiatives, as well as their approaches, goals, and challenges, thus providing a comprehensive and critical state of the art for this emerging field.

Methods

In order to map the landscape of AI initiatives in the fight against disinformation we relied on a web mapping approach (Severo and Venturini, 2016). This method operates on the idea that hyperlinks can serve as proxies for social connections. Despite the relatively low cost of creating a hyperlink, it has been consistently observed that web authors are meticulous when establishing connections. They tend to preferentially cite websites that share their thematic or social focus, and avoid citing those with opposing viewpoints, leading to picky organization of the web (Ooghe-Tabanou et al., 2018).

Websites link their discourse to other online discourses to establish hierarchies and clusters, resulting in a network of networks where densely connected zones are separated by relatively empty spaces. These territories correspond to thematic communities, where actors with similar interests and viewpoints gather. By examining the hyperlinks connecting websites dedicated to initiatives that use AI to combat disinformation, we can gain insight into the networks of actors concerned with AI counter-disinformation. In essence, knowledge of which sites are hyperlinked can reveal which actors are likely to be connected in their effort to contrast disinformation through AI.

To create a map adhering to the best practices of hyperlink analysis, we first identified a list of websites that referred to initiatives using or creating AI against disinformation. To construct the list we implemented the following steps:

1. We searched online for pre-existing lists compiled by research or public institutions.

5 https://fullfact.org/about/ai/

6 https://www.newsguardtech.com/press/launch-of-newsguard-for-ai-training-machines-with-trust-data/

7 One such failure occurred during the 2020 US presidential election when Meta employed AI tools to detect and remove false or misleading content, but these tools were not always effective, as in the case of a false claim about election fraud that spread rapidly on Facebook: https://www.nytimes.com/2020/11/23/technology/election-disinformation-facebook-twitter.html

2. We explored these lists[8,9,10,11] and found 223 websites directly related to tools, projects or initiatives of counter-disinformation.
3. We selected by manual verification 117 websites[12] of initiatives that were actively engaged[13] in the development or use of AI against disinformation.
4. We excluded inactive websites, resulting in a final list of 81 websites.

The final list of websites served as the starting point for a web crawling research. Utilizing *Hyphe* (Jacomy et al., 2016), we extracted all hyperlinks present on the websites of our list with a crawling depth of two (i.e., visiting all the pages that were two clicks away from the starting pages that we had chosen). Based on this information, we established a citation network connecting the webpages on the list. In this network, each website is represented as a node, with edges representing their incoming and outgoing hyperlinks. The final network comprises 81 nodes and 393 links. To explore its hyperlink structure, we exported and analyzed this graph on *Gephi* (Bastian et al., 2009), and, to correctly interpret and discuss the relationship which emerged within the network, we carried out a systematic reading of all the documents and media contents present on each website contained in the map focusing on the aims, approaches, and challenges highlighted by the very same initiatives.

Results

To properly read and interpret the map shown in Figures 1–3, there are several factors to consider.

Firstly, the position of nodes in space is determined by the *Force Atlas 2* algorithm, which considers the strength and type of connections between nodes.[14] The closer two nodes are in the visualization, the stronger and more numerous are their direct or indirect connections (Venturini et al., 2021).

Secondly, the heat map superimposed on the network was constructed using *Graph Recipes*.[15] This heat map shows node density, with darker gray gradients indicating higher density and lighter gray gradients representing less dense areas. The heat map is thus used to highlight the different clusters of nodes present in the network.

Third, the size of nodes and their labels are proportional to the total sum of edges entering or leaving the node, which is calculated as their degree.

Finally, node colors are also significant. In Figure 1, colors represent the modularity class of nodes as individuated by Louvain algorithms (Blondel et al., 2008). In Figure 2, colors represent geographical belonging. Finally in Figure 3, colors represent the specific category of the nodes. In this last case blue nodes represent websites related to EU-funded projects, red nodes represent research institutes (including universities and other public or private research centers), aqua blue nodes represent Information Technologies facilities (both public and private), green nodes represent fact-checking agencies, and yellow nodes represent AI tools that can be directly used to detect and counter disinformation.

Figures 1–3 allows exploring the emergence of three distinct topological areas in our network map that mostly overlap also in terms of geographical belonging and actors category.

The largest area is located at the top of the map and is dominated by European counter disinformation initiatives. This cluster is primarily composed of websites associated with Horizon 2020 projects and European research institutes. Furthermore a roughly equal number of nodes represent AI tools and IT facility sites. This latter category is exclusive to the European cluster, as IT facilities are present only in this part of the network.

The second area, located in the middle of the map, is the smallest and least densely populated of the three. This cluster serves as a transitional zone in the network (see Supplementary Table 1 for network metrics) and is primarily composed of US-based research institutes and international think tanks.

Finally, the last area is located at the bottom of the map. This cluster is characterized by the presence of the established fact-checking agencies and AI tools.

Digging deeper into the first two areas on the bottom and in the middle of the network, we can compare the approaches taken in the European Union and the United States. Firstly, the EU primarily involves large national public research centers, while the US primarily involves the academic field and actors financed by big tech. What characterizes the EU is the strict collaboration between projects like Horizon 2020 and (mostly public) IT facilities. In contrast, the peculiarity of the US area is the presence of actors financed by big tech and International institutions; like Microsoft's research institute for

8 https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html

9 https://counteringdisinformation.org/

10 https://www.weforum.org/agenda/2022/07/disinformation-ai-technology/

11 https://commission.europa.eu/strategy-and-policy/funded-projects-fight-against-disinformation

12 As a selection criterion we also look at the embeddedness of different websites. For example, if a project against disinformation contains the sites of universities participating in this project, only the project website is maintained. However, if some universities themselves participating in the project are actively engaged independently in the development or use of AI against disinformation, both websites are maintained.

13 To be engaged and thus placed on the list, an initiative must carry out one of the following three activities: (a) develop AI technology and tools; (b) lend infrastructure (e.g., computing services), or data (e.g., lists of dangerous sites), or control work (e.g., training of human-supervised algorithms); (c) systematically use the available AI tools and instruments in a counter-disinformation initiative.

14 A force vector layout works according to a physical analogy: nodes receive a repulsive force that pulls them apart, while edges act as springs that bind the nodes they connect. Once launched, the algorithm changes the layout of the nodes until an equilibrium is reached. This balance minimizes the number of line crossings and thus maximizes the readability of the graph. Not only do force vectors minimize line crossings, but they also make sense of the arrangement of nodes in space. In a network spatialized by forces spatial distance acquires meaning: two nodes are closer the more directly or indirectly connected they are (Jacomy et al., 2014). As a consequence network maps spatialized with force vectors sharply visualize clusters and connections.

15 https://medialab.sciencespo.fr/en/tools/graph-recipes/

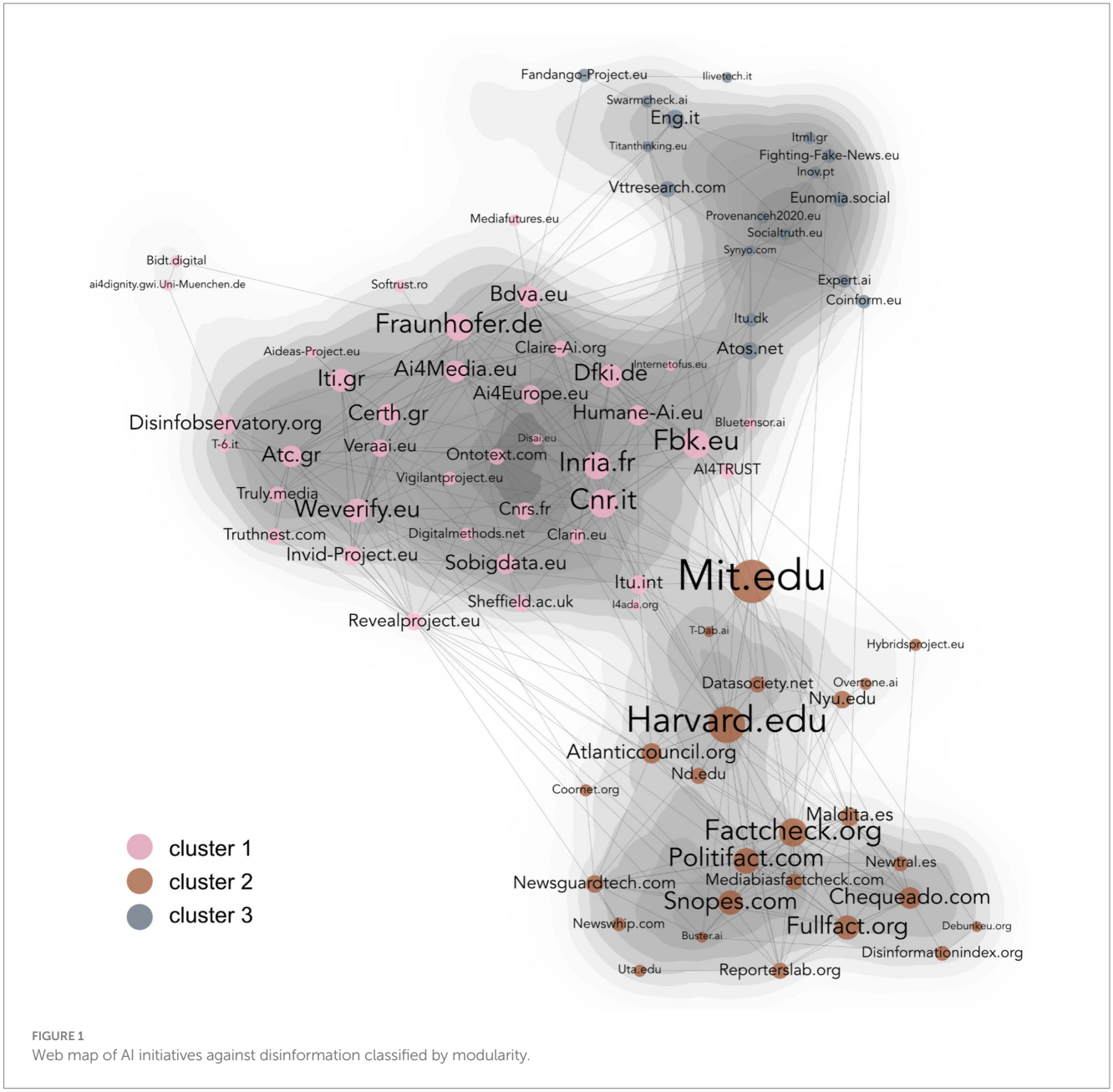

FIGURE 1
Web map of AI initiatives against disinformation classified by modularity.

Data & Society, the Atlantic Council's Digital Forensic Lab and AI for Good summit run by the International Communication Union.

Furthermore, in the EU's area, there is a majority of large national public research centers engaged in the study and development of AI to counter disinformation. In contrast, in the part of the map dominated by US institutions, it is primarily the academic field that is directly involved in these efforts.

Despite these differences, a close examination of the projects' descriptions within these two areas reveals that the strategies for developing AI tools are almost identical. In both the websites referring to EU and US areas, AI tools are described as a human-aid rather than a direct solution to the problem of disinformation while focusing primarily on improving the quality of information rather than simply targeting the spread of false information.

For example, various Horizon Europe projects are focused on building AI to navigate the vast sea of digital content and detect signals of potentially dangerous or false content. VERA.ai, building on the previous efforts of WeVerify, utilizes AI and expert crowdsourcing to detect and verify false information, including deep fakes. Projects such as AI4TRUST, REVEAL and InVID focus on developing tools to help journalists and citizens verify authentic contents. Finally, enhancing peer-to-peer moderation, SocialTruth seeks to develop blockchain solutions to counteract the online spread of disinformation, while Provenance is developing an intermediary-free solution for digital content verification that gives greater control to users of social media through AI solutions.

Similarly, in the US dominated area, initiatives such as the collaboration between NYU and Overtone.ai aim to alert readers not only about false information but also on the decontextualization of true stories to warn online users and mitigate the effects of possible negative spillovers, such as the increasing sensationalization of online debate.

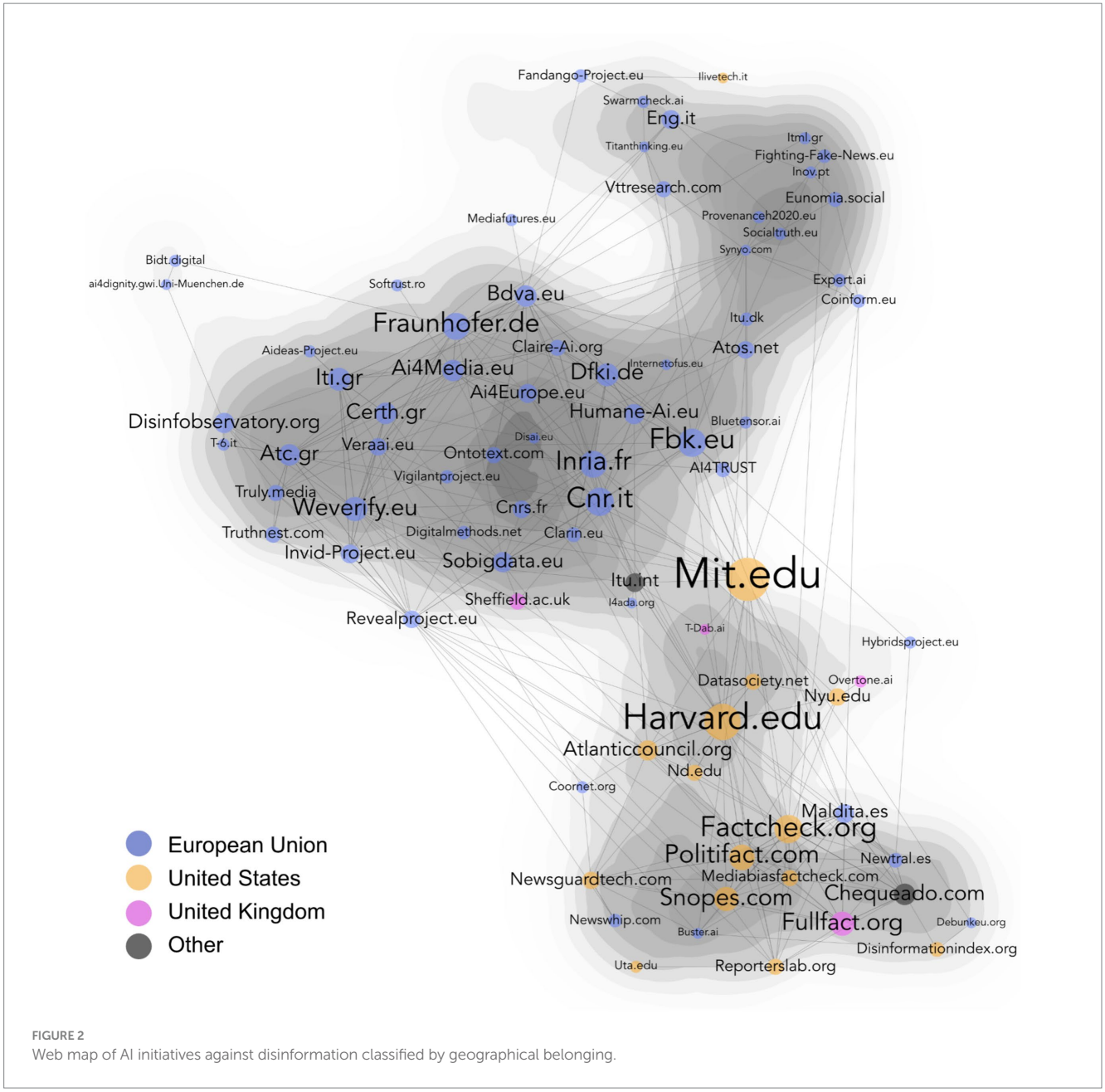

FIGURE 2
Web map of AI initiatives against disinformation classified by geographical belonging.

Overall, we can thus argue that initiatives of both the EU and US areas focus primarily on improving the quality of information through interventions on the media ecosystem, rather than simply targeting direct disinformation.

In contrast, the area located at the bottom of the map, which mainly consists of fact-checking agencies and AI tools, has a peculiar composition. This cluster is influenced both directly and indirectly by the area located in the United States that deals with legislation and policies. This is applicable not only to central U.S. based companies such as Snopes and PolitiFact but also to several initiatives such as Chequeado and FullFact, which have received funding from Google programs,[16] or as for the case of FactCheck.org that collaborated with Meta in the third-party fact-checking program of Facebook. Therefore, the development of AI initiatives against disinformation in this case is closely tied to big tech and digital platforms. Additionally, the development strategy differs significantly from the other two areas. In this sense, fact-checking agencies have a clearly different objective. These agencies use AI tools to verify news ex-post, identify false content and ultimately perform a debunking operation of incorrect information. This use of AI is mostly remedial rather than preventative.

Building on these findings, the division between the dominant approach in the fact-checking cluster and that present in the EU and US areas led us to identify two strategies. The first strategy, mainly pursued through Horizon projects in the EU and through academic research in the US, addresses the issue of disinformation 'upstream' and develops AI tools capable of improving the overall quality of information and debate in the media ecosystem. The second strategy, followed by fact-checkers, is perhaps more established and addresses the circulation of disinformation

16 https://www.poynter.org/fact-checking/2019/these-fact-checkers-won-2-million-to-implement-ai-in-their-newsrooms/

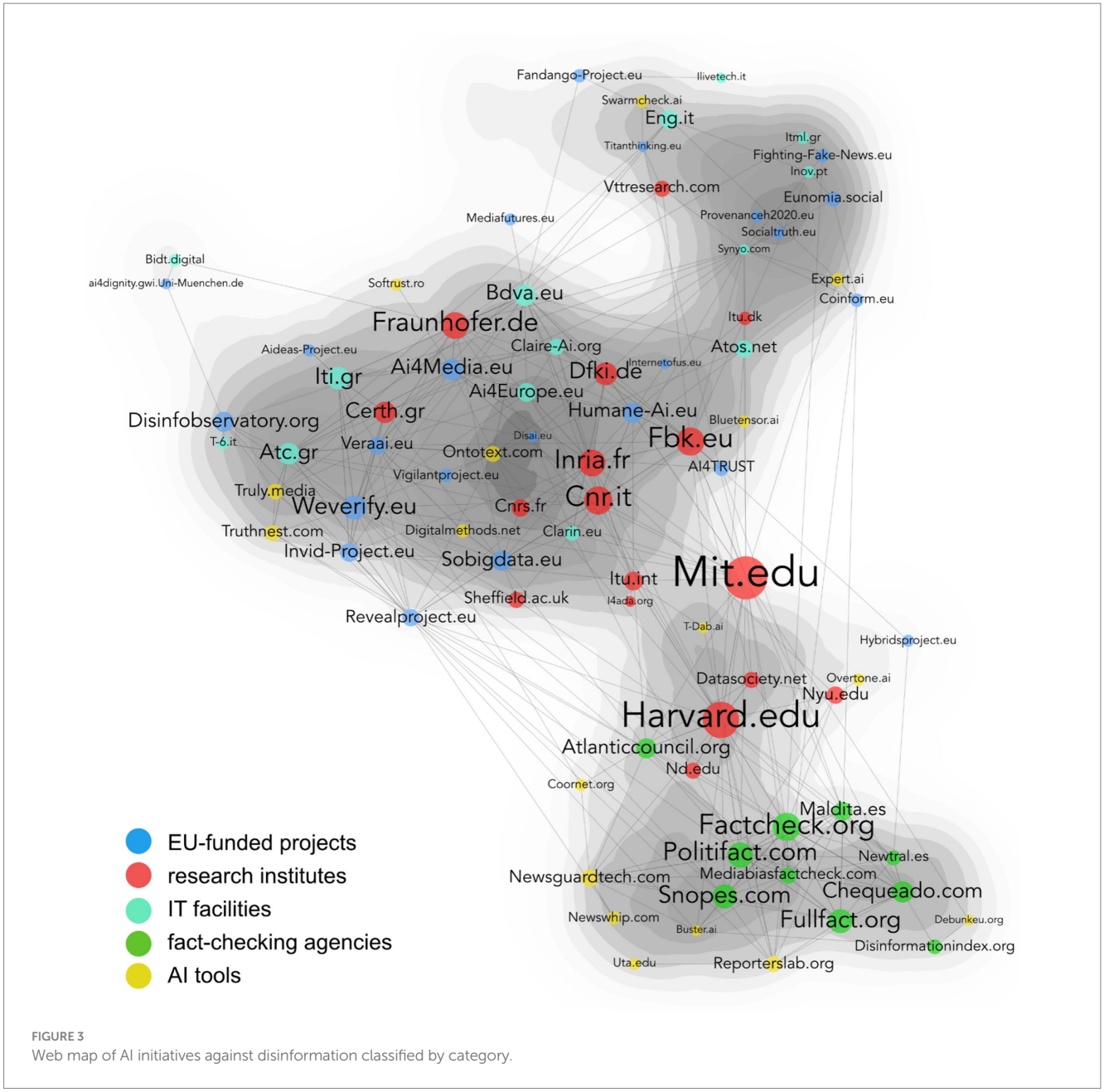

FIGURE 3
Web map of AI initiatives against disinformation classified by category.

'downstream', seeking to develop and improve the detection potential of AI.

Discussion

Our mapping illustrates how the use of AI to detect and fight disinformation is distributed in a dense network of different initiatives.

In the EU the innovation and development of AI tools is promoted by public funding, especially the H2020 program. AI tools are thus developed by European projects that are often carried out in partnership with high-education institutions, but are not *led by* them (an important exception being the University of Sheffield in the UK). In contrast, in the United States, AI tools are developed especially in higher education research environments, particularly in ivy league universities like Harvard and MIT, and supported by private funding. Both groups however have in common the strategy adopted in countering disinformation with AI, which aims at improving the overall quality of the information environment 'upstream'.

This is the crucial difference that distinguishes research projects from fact-checkers initiatives. These last initiatives, that reside primarily but not exclusively in the US, are developing and making use of AI tools to fight disinformation 'downstream', to detect disinformation narratives after their spread. Nevertheless, what is common in both those two different approaches of AI applications in the fight against disinformation is the indispensable human supervision.

Overall, the use of AI in these disinformation detection and mitigation projects presents both opportunities and challenges. On the one hand, AI can enhance the speed and accuracy of detecting and flagging potentially harmful content, allowing for faster responses to

disinformation campaigns. AI algorithms can also help identify patterns in the spread of disinformation, aiding in the development of more targeted responses. Additionally, AI can help automate certain aspects of the fact-checking process, potentially reducing the workload on human fact-checkers.

On the other hand, one major challenge is the potential for biases to be encoded into AI algorithms, which could exacerbate existing inequalities, reinforce harmful stereotypes and blur the distinction between disinformation and legitimate speech. In this sense, the issue of adversarial attacks, in which malicious actors attempt to manipulate AI systems by feeding them misleading or incorrect data, is also present.

To conclude, there are also notable gaps in our mapping effort. These gaps may indicate areas where more investigation or research may be needed. One gap is the lack of representation from non-Western countries in the network. Most of the nodes in the map are located in Europe and the United States, with only a few nodes from other regions such as the Chequeado initiative in South-America. This may be due to several factors, including limited funding and resources for AI initiatives in these regions, the issue of structural visibility in the western-driven search engines we queried, as well as different cultural and political contexts that may affect the development and implementation of AI tools to counter disinformation.

Another gap is the limited representation of civil society organizations in the network. While fact-checking agencies are represented, other types of civil society organizations such as media watchdogs and human rights groups are not as prominent. This may be because these organizations have not yet fully explored the potential of AI in their work, or because they face challenges such as limited funding and technical expertise.

Finally, it is interesting to note that most of the inactive websites excluded from the final list we investigated were projects launched in the United States after the election of Donald Trump, meaning during the height of moral panic related to perceived threats such as fake news and the post-truth era. These projects had mainly tried to solve the problem of detecting and moderating false content in a fully automated way, but they clashed with the ethical and practical limitations of such an application of AI.

Data availability statement

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

Author contributions

FP: Writing – original draft, Writing – review & editing. TV: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is part of the project "Understanding Misinformation and Science in Societal Debates" (UnMiSSeD) supported by the European Media and Information Fund.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpos.2025.1517726/full#supplementary-material

References

Alemanno, A. (2018). How to counter fake news? A taxonomy of anti-fake news approaches. *Eur. J. Risk Regul.* 9, 1–5. doi: 10.1017/err.2018.12

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In Proceedings of the international AAAI conference on web and social media. ICWSM, San Jose, California

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, P10008.

Bontcheva, K., Symeon, P., Filareti, T., Riccardo, G., et al. (2024). Generative AI and disinformation: recent advances, challenges, and opportunities". European digital media observatory. Available at: https://edmo.eu/edmo-news/new-white-paper-on-generative-ai-and-disinformation-recent-advances-challenges-and-opportunities/ (Accessed October 26, 2024).

Bontridder, N., and Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data Policy* 3:e32. doi: 10.1017/dap.2021.20

Chorąś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., et al. (2021). Advanced machine learning techniques for fake news (online disinformation) detection: a systematic mapping study. *Appl. Soft Comput.* 101:107050. doi: 10.1016/j.asoc.2020.107050

Golovchenko, Y., Hartmann, M., and Adler-Nissen, R. (2018). State, media and civil society in the information warfare over Ukraine: citizen curators of digital disinformation. *Int. Aff.* 94, 975–994. doi: 10.1093/ia/iiy148

Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of online content. *Inter. Policy Rev.* 8, 1–22. doi: 10.14763/2019.2.1407

Graves, L. (2016). Deciding what's true: the rise of political fact-checking in American journalism. New York, NY: Columbia University Press.

Graves, D. (2018). Understanding the promise and limits of automated fact-checking. Oxford: University of Oxford.

Jacomy, M., Girard, P., Ooghe-Tabanou, B., and Venturini, T. (2016). Hyphe, a curation-oriented approach to web crawling for the social sciences. In Proceedings of the international AAAI conference on web and social media. PKP Publishing Services Network: Cologne, Germany

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Force Atlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* 9:e98679. doi: 10.1371/journal.pone.0098679

Kertysova, K. (2018). Artificial intelligence and disinformation: how AI changes the way disinformation is produced, disseminated, and can be countered. *Sec. Hum. Rights* 29, 55–81. doi: 10.1163/18750230-02901005

Marsden, C., and Meyer, T. (2019). Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism. Europe: European Parliament.

Marsden, C., Meyer, T., and Brown, I. (2020). Platform values and democratic elections: how can the law regulate digital disinformation? *Comput. Law Secur. Rev.* 36:105373. doi: 10.1016/j.clsr.2019.105373

Ooghe-Tabanou, B., Jacomy, M., Girard, P., and Plique, G. (2018). Hyperlink is not dead ! In: digital tools & uses congress, Paris. New York: ACM Press.

Porter, E., and Wood, T. J. (2021). The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proc. Natl. Acad. Sci.* 118:e2104235118. doi: 10.1073/pnas.2104235118

Severo, M., and Venturini, T. (2016). Intangible cultural heritage webs: comparing national networks with digital methods. *New Media Soc.* 18, 1616–1635. doi: 10.1177/1461444814567981

Venturini, T., Jacomy, M., and Jensen, P. (2021). What do we see when we look at networks: visual network analysis, relational ambiguity, and force-directed layouts. *Big Data Soc.* 8:8488. doi: 10.1177/20539517211018488

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

Yang, K. C., and Menczer, F. (2023). Large language models can rate news outlet credibility. Available at: https://arxiv.org/abs/2304.00228 (Accessed October 26, 2024).

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 13 March 2025
DOI 10.3389/frai.2025.1474017

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Maria Chiara Caschera,
National Research Council (CNR), Italy
Liwei Yang,
Northeast Normal University, China

*CORRESPONDENCE
Yuriy Dyachenko
✉ ydyachenko@kse.org.ua

RECEIVED 31 July 2024
ACCEPTED 28 February 2025
PUBLISHED 13 March 2025

CITATION
Dyachenko Y, Humenna O, Soloviov O, Skarga-Bandurova I and Nenkov N (2025) LLM services in the management of social communications.
Front. Artif. Intell. 8:1474017.
doi: 10.3389/frai.2025.1474017

LLM services in the management of social communications

Yuriy Dyachenko[1]*, Oleksandra Humenna[2], Oleg Soloviov[3], Inna Skarga-Bandurova[4] and Nayden Nenkov[5]

[1]Kyiv School of Economics, Kyiv, Ukraine, [2]Institute of Education Content Modernization, Kyiv, Ukraine, [3]Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine, [4]Oxford Brookes University, Oxford, United Kingdom, [5]Konstantin Preslavsky University of Shumen, Shumen, Bulgaria

This paper proposes enhancing social communication management with a behavioral economics approach through artificial intelligence instruments. The research aims to explore the influence of social communication on citizens' behavior using large language model services and assess its effectiveness. The paper builds on Daniel Kahneman's dual-process theory, highlighting the intuitive system (System 1) and the rational system (System 2) in decision-making. The author introduces a third system, System 3, representing rooted in identity socially conditioned behavior influenced by societal norms and self-awareness. On this theoretical basis, the paper emphasizes automating communication management through large language model services, freeing up citizens' potential for self-determination and self-organization. By leveraging these services, messages can be crafted to support social transformation while respecting historical, cultural, and political contexts. Based on the preconditions and restrictions described above, we use GPT-4 model to generate messages based on these narratives. The experiment will use an observational study design with virtual persons. To compare the impact of original and modified messages according to the addressee's mentality, we used the Claude 3.5 Sonnet model. We can see that the potential activity of respondents after perceiving the changed message does not change much, and the original message is perceived. Modifying messages by LLM services crafted to support social transformation while respecting historical, cultural, and political contexts cause attitudes to become substantially more negative (2.5 units downward shift in median); the intentions showed a slight positive increase (0.2 units upward change in median).

KEYWORDS

dual-process theory, management of social communications, LLM services, behavioral economics, decision-making

1 Introduction

This paper proposes enhancing social communication management with a behavioral economics approach through artificial intelligence instruments.

Many studies indicate that social communications undergo significant changes under the influence of digital technologies and social media (Halich et al., 2023; Ming and Salman, 2023). Social media increasingly play a crucial role in cultural and social life, shaping public opinion and stimulating discussions on important societal issues. Social communications also impact the development of intellectual processes in group situations, promoting the formation of "collective intelligence." This occurs through adaptive communication networks that can change and restructure according to context, enhancing the efficiency of group decisions. Scientists note that such networks allow groups to exchange information better and compare

different viewpoints, promoting optimal information flow and improving the quality of decisions made.

Social communications have dramatically transformed in the digital age, with social media platforms increasingly shaping public opinion, cultural discourse, and collective decision-making. These digital networks have created new "collective intelligence" forms through adaptive communication systems that allow groups to exchange information, compare perspectives, and make decisions more efficiently. However, this evolution in social communications presents opportunities and challenges for effectively managing public discourse and understanding its influence on citizen behavior.

The complexity of human decision-making in this social communications landscape can be understood through Daniel Kahneman's dual-process theory, which describes two cognitive systems: the fast, intuitive System 1 and the slower, rational System 2. Building on this framework, the research proposes a third system – System 3 – which accounts for socially conditioned behavior shaped by cultural norms, identity, and collective values. This triadic model provides a more comprehensive framework for understanding how individuals make decisions within their social and cultural context, particularly in digital environments where social influence is increasingly pervasive.

The research proposes leveraging behavioral economics principles in conjunction with large language model (LLM) services to address the challenges of managing social communications in this complex landscape. This innovative approach aims to enhance the effectiveness of social communication management by accounting for all three cognitive systems – intuitive responses, rational deliberation, and socially conditioned behavior. By incorporating these advanced AI tools while considering the multifaceted nature of human decision-making, the research seeks to develop more nuanced and effective strategies for understanding and influencing citizen behavior through social communications.

Daniel Kahneman's dual-process theory presents a fascinating framework for understanding human decision-making through two distinct cognitive systems: System 1 and System 2 (Kahneman, 2011). This model illuminates the dynamic interplay between intuition and rationality in our cognitive processes, impacting everything from mundane daily choices to critical life decisions.

System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. This system is often referred to as the intuitive system because it involves immediate, gut-response decision-making that does not require conscious thought. The operations of System 1 are typically fast, automatic, unconscious, and rely on heuristic patterns.

Conversely, System 2, the rational system, involves more deliberate, effortful, and conscious decision-making processes. It allocates attention to effortful mental activities that demand it, including complex computations and formulating reasoned arguments. System 2 is slower and more methodical, often engaging in a critical evaluation of the outcomes generated by System 1. This system is typically activated when a person needs to focus on a task that requires logical reasoning or when making decisions that require careful consideration, such as calculating a math problem or deciding on a moral dilemma.

The interplay between these two systems is crucial for understanding human behavior. System 1's automatic operations can sometimes lead to biases and errors in judgment due to its reliance on associative memory and heuristic thinking. However, it is often efficient and effective for routine decision-making.

This dual-process approach to decision-making suggests that while our behavior may initially be guided by System 1's fast and automatic responses, it is often overseen and corrected by System 2's reflective capabilities. System 2 ensures that our actions align with broader cognitive evaluations and moral judgments. It acts as a monitor and a control mechanism that checks and occasionally corrects the impressions and decisions suggested by System 1.

The proposed work builds on seminal work in dual-process theory – for example, Kahneman's (2011) distinction between fast, intuitive (System 1) and slow, deliberate (System 2) decision-making – and extends this framework by introducing a third system (System 3) that captures the influence of socially rooted, identity-driven behavior. This extension is motivated by critiques that binary models can be overly reductive when applied to complex social communication (e.g., Evans and Stanovich, 2013). Recent reviews have proposed integrating insights from embodied and predictive processing into dual-process models, arguing that System 1 processes are automatic and shaped by embodied social experiences (Bellini-Leite, 2022).

Comparative studies in behavioral economics have applied dual-process and nudging theories to practical interventions. For instance, randomized controlled trials have shown that well-designed nudges can reliably change health or safety behaviors by altering environmental cues. At the same time, emerging research on AI in behavioral economics has demonstrated that LLMs can be harnessed to automate persuasive communication. Rahwan and colleagues have conceptualized AI systems as social actors influencing trust and cooperation through human-like interaction patterns (Rahwan et al., 2019). Some perspectives, such as those articulated by Camerer on the interaction of AI with human decision-making, further support the promise of these technologies in complex social settings (Camerer, 2018).

The synthesis of these varied approaches suggests that while traditional social communication management has long relied on human-mediated control to safeguard authenticity and self-determination, recent advances in AI offer promising avenues for scalable, personalized interventions. By introducing System 3, the current proposal seeks to integrate dual-process insights with social identity and norm-based theories, accounting for culturally and historically situated communication practices. Nevertheless, several challenges remain. In particular, the operationalization of System 3 is still under debate, and conflicts persist between the conventional wisdom favoring autonomous citizen engagement and the risks of algorithmic manipulation. The comparative Table 1 summarizes the theoretical underpinnings, methodologies, and limitations of related works in this emerging field.

2 Materials and methods

2.1 Enhancing the dual-process theory of decision-making

Building on Daniel Kahneman's dual-process theory of human cognition, which identifies System 1 (intuitive) and System 2 (rational) as key frameworks in decision-making, it is beneficial to consider introducing a third system, System 3, to capture a broader spectrum

TABLE 1 Comparative table of related works.

Study/approach	Theoretical framework	Methodology	Key Findings	Limitations
Kahneman's dual-process theory	Fast (System 1) vs. slow (System 2) decision-making (Kahneman, 2011); extended by Evans and Stanovich (2013)	Reviews of experimental studies on judgment and decision-making	Established robust evidence for two distinct cognitive processes	Criticized for oversimplifying the nuanced effects of social identity and contextual influences
Embodied and predictive extensions	Integration of embodied cognition with dual-process models (Bellini-Leite, 2022)	Meta-analyses and conceptual reviews using neuroimaging and behavioral experiments	Suggest that automatic responses (System 1) are shaped by embodied experiences, urging the addition of further processing layers	Ongoing debate regarding the operational definition and empirical separability of the proposed third system (System 3)
Nudging interventions in behavioral economics	Libertarian paternalism and choice architecture (Thaler and Sunstein, 2008)	Randomized controlled trials (RCTs) in health, finance, and safety interventions	Demonstrated that subtle changes in choice architecture can effectively shift behavior	Effectiveness varies with context and individual differences; ethical concerns regarding manipulation
AI empowerment in behavioral sciences	Integration of behavioral economics with AI tools (Thaler and Sunstein, 2008)	Field experiments and case studies using LLMs for message generation	Personalized AI interventions can enhance decision-making efficiency and support social communication	Challenges remain in scaling interventions and mitigating algorithmic biases; often reliant on proprietary platforms
Machine behavior and social actor models	Conceptualizing AI systems as social actors using computational social science frameworks (Rahwan et al., 2019)	Mixed-method approaches, including observational studies and experimental designs	AI systems display human-like interaction patterns that can influence trust and cooperative behavior	Debate continues whether AI can fully replicate human social behavior; ethical and transparency issues persist
Proposed approach (enhancing social communication management)	Extension of dual-process theory by adding System 3 to account for socially conditioned, identity-driven behavior	Observational study design with virtual persons using LLMs for automated message generation	Suggests that automated communication management can free citizens' potential for self-determination while enabling social transformation	Novelty of System 3 definition; observational design limits causal inference; potential conflicts with traditional views on citizen autonomy

of human behavior foundations. This proposed System 3 would represent socially conditioned behavior influenced by societal norms and self-awareness, encapsulating the intricate ways in which our decisions are shaped not only by subconscious and rational factors but also by our sociocultural environment.

Proposed System 3 reflects the influencing of social conditioning, identity, and societal norms on behavior and decision-making, which were investigated within the framework of Social Identity Theory (Tajfel and Turner, 1979), explains how people's sense of self is derived from group memberships and social categories, which then influence behavior and decision-making. From this point of view, a person who strongly identifies as environmentally conscious may automatically choose eco-friendly products without engaging in the cost–benefit analysis typical of System 2 thinking, yet this is not quite the automatic, instinctive processing of System 1 either.

Within the Cultural-Historical Activity Theory (Vygotsky, 1978; Leontiev, 1978), cultural and historical contexts shape human behavior and cognition through internalized social practices. From this point of view, for example, the way people queue in different cultures – British people tend to form orderly lines automatically, while other cultures might have different spatial arrangements for waiting that are neither purely instinctive (System 1) nor calculated (System 2).

Habitus Theory (Bourdieu, 1977) explains how social conditioning becomes embodied in individuals, creating durable dispositions that guide behavior without conscious calculation. For example, class-based differences in food preferences or cultural tastes that feel "natural" to individuals but are socially conditioned.

Socially conditioned behaviors influenced by identity and norms are very real. System 3 can integrate its aspects into the decision-making framework to consider the impact of social and cultural influences on cognition and behavior.

System 3 may reflect a more reasoned decision-making process in which inputs from sensory experiences are integrated and analyzed based on an individual's values and goals. Furthermore, System 3's sensitivity to feedback loops based on human values implies that our reflective processes are logical and influenced by our personal and societal norms.

System 3 can be understood as a very slowly changing socially conditioned system rooted in identity that plays a pivotal role in behaviors driven by societal norms and expectations. It involves an awareness of societal values and the reflexive ability to adjust individual behavior following these values, which can be described as "decent" or socially responsible behavior. For instance, when individuals decide to follow recycling guidelines or adhere to public health advisories, they are likely influenced by System 3, which mediates personal actions with a collective ethical framework.

This introduction provides an opportunity to integrate the cognitive influences of Systems 1 and 2 with the external social

context, creating a triadic model of human cognition in which behavior is also a function of social conditioning and identity. System 3 is slower in its evolution, reflecting the gradual changes in cultural norms and societal values over time. It provides a feedback mechanism incorporating societal approval or disapproval into personal decision-making processes, influencing long-term behavioral patterns and ethical considerations.

Moreover, System 3 enriches our understanding of identity as it interacts with subconscious instincts (System 1) and rational considerations (System 2). Identity in this context is not merely about personal introspection or rational self-assessment but also involves how individuals see themselves in the social mirror. This aspect of self-awareness is shaped by cultural, social, and historical forces, and it affects how individuals conform to or rebel against societal expectations.

Incorporating System 3 into the dual-process framework may allow for a more nuanced understanding of human behavior, including the influence of socially determined norms. This triadic model explains individual decision-making and enhances our comprehension of group behaviors and societal trends. It accounts for the complexity of decisions not entirely based on instinct or reason but also deeply embedded in a social matrix that dictates what is considered appropriate, responsible, or ethical.

Thus, for a comprehensive analysis of human behavior, it is essential to consider all three systems: the subconscious instincts governed by System 1, the rational deliberations of System 2, and the socially conditioned behaviors of System 3. Each system interacts with the others, creating a dynamic interplay that profoundly shapes individual and collective behavior. By acknowledging the role of societal norms and identity (System 3), we gain a fuller picture of the factors that drive human actions and the societal frameworks that guide them.

2.2 Application of the enhanced dual-process theory of decision-making for social communication management

Social communication management represents a critical dimension of understanding human interactions within various contexts, particularly considering the extended framework of human cognition involving Systems 1, 2, and 3 – where System 3 specifically deals with socially conditioned behavior.

System 1, the intuitive system, is implicit in social communication by facilitating quick, automatic responses to social stimuli. This system allows individuals to respond to social cues quickly and efficiently, crucial in dynamic social interactions.

System 2, the rational system, introduces a more deliberate layer to social communication. It governs how we process complex linguistic constructs, interpret semantic ambiguities, and engage in thoughtful discourse.

System 3, based on identity, as introduced, encompasses the socially conditioned behaviors heavily influenced by societal norms and cultural expectations. This system is pivotal in managing social communication because it guides individuals on appropriate social behaviors and communication ethics (Maia, 2014). It involves a higher level of self-awareness and reflection (Dyachenko et al., 2018), allowing individuals to adjust their communicative practices to align with social norms and values.

Effective social communication management, therefore, requires an interplay of all three systems. The intuitive reactions from System 1 need to be balanced with the thoughtful analysis of System 2 and the ethical considerations of System 3. This balance is crucial in diverse social contexts, such as multicultural interactions, where miscommunications can arise from varying social norms and expectations.

However, managing this balance presents challenges. Over-reliance on System 1 can lead to premature judgments or cultural faux pas, whereas excessive deliberation in System 2 can result in communication that is perceived as insincere or overly calculated. Moreover, System 3's slow adaptability might lag rapidly changing societal norms, leading to outdated or inappropriate responses in new social milieus.

The ability to comprehend and engage with a message is influenced by intelligence, available time, knowledge level, environmental distractions, and message repetition frequency. People are more likely to respond to messages that align with their existing knowledge (Ming and Salman, 2023).

Therefore, social communication management involves understanding and integrating cognitive processes across all three systems. By fostering awareness of these systems and their impact on communication, individuals and organizations can enhance their interactions within and across cultural boundaries, leading to more effective and harmonious social relations. This approach improves interpersonal communications and equips society to better navigate the challenges of an increasingly interconnected world.

2.3 Artificial intelligence instruments for communication management

Integrating AI instruments (Zerfass et al., 2020), particularly LLMs, into communication management (Seidenglanz and Baier, 2023) represents a significant advancement in technology and social interaction. We can use AI-driven tools to enhance communication strategies by utilizing the cognitive systems framework: System 1 (intuitive), System 2 (rational), and System 3 (socially conditioned), as discussed previously.

LLMs like GPT (Generative Pre-trained Transformer) have revolutionized the field of natural language processing (NLP) by enabling machines to understand and generate human-like text. These models operate at the intersection of all three cognitive systems. They mimic System 1 by rapidly generating responses based on patterns learned from vast datasets, facilitating quick and intuitive communication. In System 2, LLMs assist in processing complex information and producing reasoned, coherent outputs necessary for detailed explanations or problem-solving tasks in communications. Finally, through learned social and cultural nuances, LLMs reflect aspects of System 3 by adhering to societal norms and ethical guidelines in their outputs. This is crucial for maintaining decency and appropriateness in communication.

In content creation, AI tools are adept at generating written content for blogs, reports, and marketing materials. They assist in crafting messages that are not only grammatically correct and stylistically appropriate but also tailored to the cultural context of the target audience, showcasing their alignment with System 3.

In social media management, LLMs analyze and generate content for social media platforms, managing posts and interactions to

conform to social norms. They can moderate discussions, filter inappropriate content, and maintain a brand's voice across channels, demonstrating an advanced application of System 3 in communication.

While LLMs offer substantial benefits, their application in communication management is not without challenges. One major concern is the potential for these models to propagate biases in their training data, which can lead to inappropriate or harmful communication outputs. This issue ties directly into the ethical considerations of System 3, where societal norms are paramount – ensuring that LLM outputs are unbiased and representative requires continuous monitoring and updating of AI models to reflect evolving societal values.

Another challenge is the risk of dependency on automated systems, which might degrade human cognitive abilities in Systems 2 and 3. Relying too heavily on AI for communication tasks might diminish individuals' ability to engage deeply with complex information or navigate social nuances independently.

Thus, applying LLMs in communication management demonstrates a powerful synergy between AI technology and human cognitive frameworks. By enhancing intuitive, rational, and socially conditioned communication processes, LLMs improve the efficiency and effectiveness of communication strategies and present new opportunities and challenges in the digital age. As these tools become more integrated into societal frameworks, responsibly managing their development and usage will be crucial to maximizing their benefits while mitigating risks. This will ensure that AI-enhanced communication supports broader equity goals, ethical interaction, and cultural sensitivity.

2.4 Application of LLM services for communication management to free up citizens' potential for democratic transformations

Integrating LLMs in communication management can significantly influence democratic processes by freeing up citizens' potential for participation and engagement. LLM services can enhance democratic transformations by facilitating informed discourse and engagement, aligned with the cognitive systems framework: System 1 (intuitive), System 2 (rational), and System 3 (socially conditioned).

Through their capacity to process and generate vast amounts of information quickly, LLMs can democratize access to complex data and policy discussions. By simplifying intricate government documents, legal texts, and policy debates into more accessible language, LLMs can empower citizens by making information more understandable and engaging. This aligns with System 1, providing intuitive grasps of complex subjects without overwhelming cognitive load, thus encouraging broader public participation in democratic discourse.

Regarding System 2 operations, LLMs can contribute to more rational and informed public discussions by providing data-driven insights and balanced perspectives. For instance, during election periods, LLMs can analyze candidates' proposals, offering unbiased summaries and comparisons based on historical data and policy analysis. This rational, evidence-based approach to communication helps counter misinformation and promote critical thinking among the electorate, which is essential for informed voting and civic participation.

Reflecting System 3, LLMs can be pivotal in maintaining and promoting democratic norms and social cohesion. By moderating online forums and social media platforms, LLMs can help enforce community guidelines that discourage hate speech and encourage respectful discourse. Additionally, LLMs can facilitate cross-cultural and inter-community dialogues by translating content and mediating discussions, fostering inclusivity and mutual respect among diverse groups.

However, deploying LLMs in democratic contexts must be navigated carefully to avoid potential pitfalls. The risk of perpetuating existing biases, manipulating opinions, or infringing on privacy remains significant. Ensuring that LLMs operate transparently and ethically is paramount to maintaining trust in democratic institutions and processes. Continuous monitoring and adaptive frameworks are necessary to align LLM outputs with evolving societal values and democratic principles.

Moreover, the dependency on technology for democratic engagement could lead to disparities in participation among different demographic groups, especially those with limited access to digital technologies or those less technologically literate. Addressing these disparities is crucial to ensure that the benefits of LLMs in democratic transformations are equitably distributed.

Applying LLMs in communication management holds considerable potential to enhance democratic transformations by making information more accessible, supporting rational discourse, and reinforcing social norms conducive to democracy. By effectively integrating these systems into democratic processes, LLMs can help free up citizens' potential for active and informed participation. However, managing these technologies to ensure they uphold democratic values without compromising ethical standards or social equity will be critical in realizing their full potential in supporting democratic transformations. This approach not only enhances the immediate efficiency of democratic engagements but also contributes to the long-term resilience and inclusivity of democratic institutions.

Modification messages and receiving feedback through simulated virtual personalities through LLM services are two ways to manage the impact on audience attitudes and behavioral intentions. LLMs, such as OpenAI's GPT series, can generate human-like text, enabling the tailoring of messages to specific audiences. This customization can enhance message effectiveness by aligning content with audience preferences and expectations. However, the ethical implications of such modifications, including concerns about authenticity and manipulation, warrant careful consideration.

Recent studies have explored the impact of virtual personalities in LLM-driven communications. Hu and Collier (2024) investigated how incorporating persona variables – demographic, social, and behavioral factors – affects LLMs' ability to simulate diverse perspectives. Their findings suggest that while persona – integration offers modest improvements in simulation accuracy, the overall effect size is limited, indicating that persona variables account for less than 10% of the variance in human annotations.

The effectiveness of personality-adaptive chatbots has also been systematically reviewed; Ait Baha et al. (2023) analyzed 66 studies focused on chatbots that adapt to users' personalities, highlighting various deep learning approaches for personality recognition and adaptation. The review emphasizes aligning chatbot responses with user personalities to enhance engagement and satisfaction. However, challenges remain in accurately recognizing and adapting to diverse personality traits.

Furthermore, the ability of LLMs to emulate human personalities has been examined. Li et al. (2024) demonstrated that supervised fine-tuning can more effectively shape LLM personalities than prompt-based approaches, resulting in higher validity and consistency in personality emulation.

In this work, we propose a low-requirements technique using LLM prompts to modify messages targeting specific behaviors or attitudes by tone, framing, and content. We used GPT-4 model from OpenAI to generate messages and Claude 3.5 Sonnet model from Anthropic to access an LLM and compare the impact of original and modified messages. This technique explores different aspects of identity and worldview alignment. The effectiveness will be estimated by the impact on the behavior of virtual personalities created through the LLM service (Figure 1).

3 Results

3.1 Qualitative and quantitative analysis of the effectiveness of the impact of artificial intelligence instruments on social communication

Practically, we emphasize automating communication management through LLM services, freeing up citizens' potential for self-determination and self-organization. By leveraging these services, messages can be crafted to support social transformation while respecting historical, cultural, and political contexts.

We can design narratives in support of the transformation of the country with a deep understanding of the historical, cultural, and political contexts. The narratives should be respectful, informed, and considerate of the diverse perspectives within the country. We can propose the following narrative topics that could be explored, considering the need for sensitivity and respect for historical diversity, decentralization of power, economic opportunities, cultural renaissance, peace and cooperation, environmental management, and global integration.

Based on the preconditions and restrictions described above, we used GPT-4 model from OpenAI to generate messages on these narratives.

Designing an experiment to assess the effectiveness of LLM-generated messages without direct feedback from participants requires indirect evaluation methods. Such an experiment would rely on observable behaviors and data analysis techniques to infer the effectiveness of communication based on behavioral economics principles and cognitive responses. Here's a proposed experiment design overview.

The objective is to evaluate the effectiveness of messages generated by LLMs in terms of identity resonance, worldview alignment, and loyalty induction using observable behaviors of virtual persons.

Using virtual personalities through LLM services for survey simulations gives researchers a powerful tool to explore diverse response patterns in a controlled environment. By crafting and deploying distinct personas with varying backgrounds, attitudes, and preferences, researchers can simulate various respondent types without relying exclusively on real-world participants. This approach

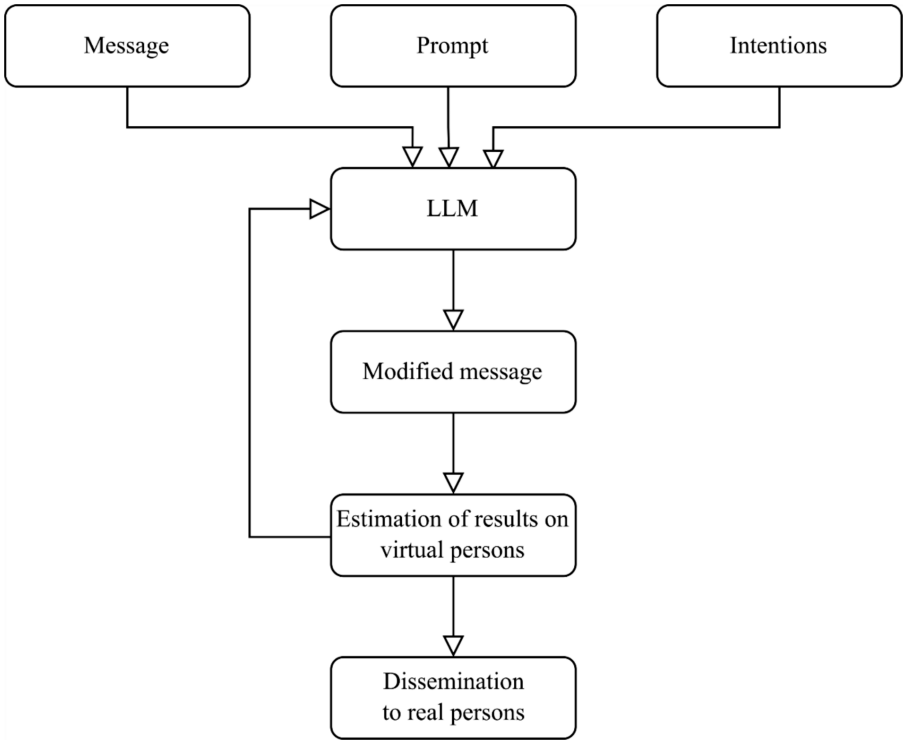

FIGURE 1
Scheme of proposed modification of messages by LLM.

broadens the scope of data collection and addresses ethical and logistical concerns, as it minimizes the need to involve human subjects in repetitive or sensitive questioning.

The robustness of this approach arises from the ability to test and validate findings against multiple simulated perspectives. Instead of relying on a single, uniform data source, researchers can compare and contrast results across various personas, each serving as an independent test case. This triangulation of evidence bolsters the credibility of any conclusions drawn, ensuring that the observed patterns are not merely artifacts of one particular group or sample. Consequently, using virtual personalities in LLM-based survey simulations offers a scalable, ethically sound, and methodologically robust means of generating and testing hypotheses in various research fields.

We propose the following experimental design:

1. Message creation. We use an LLM to generate multiple versions of messages targeting specific behaviors or attitudes. These messages should vary systematically in tone, framing, and content to explore different aspects of identity and worldview alignment. Alongside LLM-generated messages, we create control messages using standard communication practices.
2. Behavioral modelling. To take engagement metrics, we can track user interactions with each message, such as likes, shares, comments, and time spent on message pages.
3. Data analysis. Comparing engagement and behavior metrics between messages to determine which versions most effectively influence behavior suggests higher effectiveness regarding identity resonance and worldview alignment.

As the key variables, we define the following:

1. Independent variable – the message type (LLM-generated vs. control).
2. Dependent variable – user engagement.

To provide data privacy, we ensure all data collection complies with privacy laws and platform policies, using only anonymized, aggregated data for analysis. While direct feedback is not collected, ensure users know general data usage policies on platforms.

As a limitation, we see:

1. Interpretation bias that, without direct feedback, requires consideration that must be made about why users engage differently with various messages.
2. External factors like current events or social trends could affect user behavior independently of the messages.

This experiment design leverages indirect measures to assess the impact of LLM-generated messages on user behavior, providing insights into their effectiveness in communication management. By analyzing engagement in response to different message strategies, one can infer how well messages resonate with user identities and worldviews and how effectively they induce loyalty without direct user feedback.

Concerning difficulties in disseminating and reviewing feedback on messages, we propose using virtual persons.

We used the Claude 3.5 Sonnet model from Anthropic to access an LLM and compare the impact of original and modified messages according to the addressee's mentality.

Argyle et al. (2023) shows that the GPT-family language model is "both fine-grained and demographically correlated, meaning that proper conditioning will cause it to emulate response distributions from various human subgroups accurately. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and socio-cultural context that characterize human attitudes".

But in Dillion et al. (2023) outlined caveats of using AI as a participant: "LLMs may be most useful as participants when studying specific topics, when using specific tasks, at specific research stages, and when simulating specific samples".

We created 10 virtual personalities and set the task of modeling their response to the original and modified information messages.

We evaluated the responses on two scales:

1. Attitude to the message on a scale from −3 "extremely negative" to +3 "extremely negative."
2. Intention to act due to the influence of the information received on a scale from 0, "I am not going to do anything at all," to 3, "I will definitely take certain actions."

The graphical interpretation of the estimates of virtual respondents' answers is shown in Figure 2.

4 Discussion

We can see that the potential activity of respondents after perceiving the changed message does not change much, and the original message is perceived.

As a result of changing the message by LLM service, we got the following results:

1. Respondents' attitudes became substantially more negative (2.5 units downward shift in median).
2. The intentions showed a slight positive increase (0.2 units upward shift in median).
3. There is a decoupling between attitudes and intentions in the modified message condition.

Generally, messages crafted to support social transformation while respecting historical, cultural, and political contexts tend to be more pessimistic about events and significantly impact the intention to act.

The growing willingness to act under the influence of modified messages can free up citizens' potential for self-determination and self-organization.

However, in the future, we must consider ways to reduce the decoupling between attitudes and intentions in the modified message condition.

One of the primary difficulties in using virtual personalities for survey simulations lies in ensuring the authenticity and representativeness of the generated responses. Although LLMs can capture various conversational styles and viewpoints, these may not perfectly mirror the diversity of real human populations. In some cases, the pretraining data for LLMs might be skewed toward specific

10.3389/frai.2025.1474017

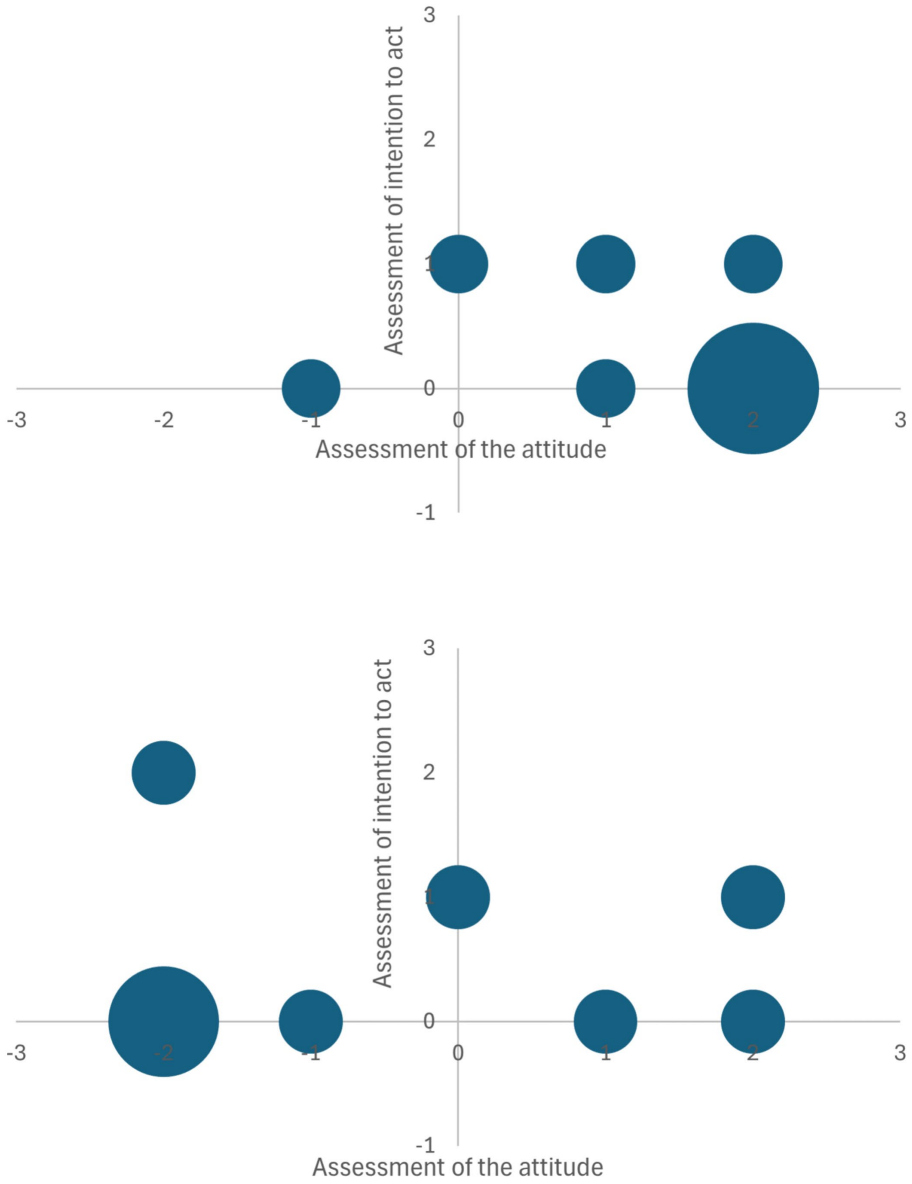

FIGURE 2
Graphical interpretation of the estimates of virtual respondents' response to the original (top) and modified information messages (bottom).

demographics or cultural contexts, limiting the models' capacity to simulate particular population segments reliably. Overcoming this hurdle involves curating high-quality, balanced datasets and continuously refining the underlying models to incorporate a broad spectrum of cultural, geographical, and socio-economic factors. Researchers may also need to apply domain-specific fine-tuning so that the virtual personalities accurately reflect the nuances of the populations they aim to represent.

Another challenge arises in managing and mitigating the risk of unintended biases or inconsistencies introduced by virtual personas. Since LLMs can inadvertently reproduce prejudices embedded in their training data, certain personality constructs may display biased or conflicting views across different topics. To address this issue, researchers can employ systematic bias-detection protocols – such as leveraging third-party tools or custom scripts to analyze generated text for indicators of stereotyping or discrimination. Implementing iterative feedback loops, wherein researchers test simulated survey results against smaller-scale real-world samples, can help ensure that the virtual personalities remain as bias-free and consistent as possible.

While these advancements offer promising avenues for enhancing communication through LLMs and virtual personalities, it is crucial to address ethical considerations. The potential for manipulation and the authenticity of AI-generated content pose significant challenges. Ongoing research is essential to develop guidelines and frameworks that ensure the responsible use of LLMs in modifying messages and simulating virtual personalities, thereby safeguarding against ethical pitfalls and promoting trust in AI-mediated communications.

Due to significant ethical challenges, particularly concerning biases and manipulation in integrating AI into social communication management, for instance, gender and racial biases, reinforcing stereotypes, and marginalizing certain groups (Rallabandi et al., 2023), several strategies can mitigate these ethical concerns. We must

incorporate diverse and representative data during the training phase to minimize inherent biases encompassing a wide range of demographics and perspectives, which can lead to more equitable AI outcomes. Another effective strategy is the interdisciplinarity through establishing teams in AI development with ethicists, sociologists, and domain experts alongside engineers, organizations who can better anticipate and address ethical dilemmas related to AI deployment in social communication that fosters a more holistic understanding of potential impacts and promotes the creation of systems that align with societal values. Regular audits and assessments of AI systems by conducting periodic evaluations to detect and rectify biases and ensure that AI tools remain fair and effective over time are also essential. Adhering to established ethical guidelines and frameworks, such as The Toronto Declaration (2018) is vital in guiding the responsible use of AI in social communication management to emphasize the protection of human rights in machine learning systems, advocating for accountability and the mitigation of discrimination.

5 Conclusion

The research reveals insights into leveraging large language model (LLM) services for social communication management by introducing an innovative triadic cognitive framework that extends Kahneman's dual-process theory. By proposing System 3 – a socially conditioned behavioral system rooted in identity and cultural norms – the study provides a more nuanced understanding of human decision-making processes. Integrating AI technologies with this enhanced cognitive model demonstrates the potential for more sophisticated communication strategies that respect historical, cultural, and political contexts while facilitating democratic transformations.

Experimental results using virtual personalities highlight the promise and challenges of AI-driven communication management. While LLM-modified messages showed a significant negative shift in attitudes (2.5 units downward), they paradoxically generated a slight positive increase in behavioral intentions (0.2 units upward). This unexpected outcome underscores the complex interactions between message framing, cognitive systems, and potential behavioral responses. The research emphasizes the need for careful, ethically-guided approaches to AI-mediated communication that avoid manipulation while supporting informed civic engagement.

Looking forward, this research opens critical avenues for future exploration in AI-enhanced social communication. Key challenges remain in mitigating potential biases, ensuring authentic representation, and developing robust frameworks for responsible AI deployment. Interdisciplinary collaboration involving technologists, ethicists, sociologists, and communication experts will be crucial in refining these approaches. By continuing to develop nuanced, context-sensitive AI communication tools that respect individual and collective cognitive processes, we can potentially unlock new pathways for more inclusive, informed, and transformative social interactions.

Data availability statement

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

Author contributions

YD: Conceptualization, Methodology, Project administration, Writing – original draft, Writing – review & editing. OH: Writing – review & editing. OS: Writing – review & editing. IS-B: Writing – review & editing. NN: Writing – review & editing.

Funding

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

Acknowledgments

In the paper used written and visual content produced by GPT-4 and Claude 3.5 Sonnet models.

Conflict of interest

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

Publisher's note

References

Ait Baha, T., El Hajji, M., Es-Saady, Y., and Fadili, H. (2023). The power of personalization: a systematic review of personality-adaptive Chatbots. *SN Comput. Sci* 4:661. doi: 10.1007/s42979-023-02092-6

Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Polit. Anal.* 31, 337–351. doi: 10.1017/pan.2023.2

Bellini-Leite, S. C. (2022). Dual process theory: embodied and predictive; symbolic and classical. *Front. Psychol.* 13:805386. doi: 10.3389/fpsyg.2022.805386

Bourdieu, P. (1977). Outline of a theory of practice, vol. *16*. New York: Cambridge University Press.

Camerer, C. F. (2018). "Artificial intelligence and behavioral economics" in The economics of artificial intelligence: An agenda, vol. *2018* (Chicago, IL, USA: Univ. Chicago Press), 587–608.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends Cogn. Sci.* 27, 597–600. doi: 10.1016/j.tics.2023.04.008

Dyachenko, Y., Nenkov, N., Petrova, M., Skarga-Bandurova, I., and Soloviov, O. (2018). Approaches to cognitive architecture of autonomous intelligent agent. *Biol. Inspir. Cogn. Architect.* 26, 130–135. doi: 10.1016/j.bica.2018.10.004

Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Halich, A., Kutsevska, O., Korchagina, O., Kravchenko, O., and Fiedotova, N. (2023). The influence of social communications on the formation of public opinion of citizens during the war. *Soc. Legal Stud.* 6, 43–51. doi: 10.32518/sals3.2023.43

Hu, T., and Collier, N. (2024). Quantifying the persona effect in LLM simulations. In Proceedings of the 62nd annual meeting of the Association for Computational Linguistics. Vol. 1 10289–10307.

Kahneman, D. (2011). Thinking, fast and slow. London: Penguin Books.

Leontiev, A. N. (1978). Activity, consciousness, and personality. New Jersey: Prentice-Hall Englewood Cliffs.

Li, W., Liu, J., Liu, A., Zhou, X., Diab, M., and Sap, M. (2024). BIG5-CHAT: shaping LLM personalities through training on human-grounded data. Available at: https://arxiv.org/abs/2410.16491

Maia, R. C. M. (ed.). (2014). "Mass media representation, identity-building and social conflicts: towards a recognition-theoretical approach" in Recognition and the media (London: Palgrave Macmillan).

Ming, Y., and Salman, A. (2023). A study in the influence of social media on communicative behavior in all walks of life. *Int. J. Educ. Psychol. Counsel.* 8, 297–309. doi: 10.35631/IJEPC.852023

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., et al. (2019). Machine behaviour. *Nature* 568, 477–486. doi: 10.1038/s41586-019-1138-y

Rallabandi, S., Kakodkar, I. G. S., and Avuku, O. (2023). Ethical use of AI in social media. 2023 International Workshop on Intelligent Systems (IWIS), 1–9.

Seidenglanz, R., and Baier, M. (2023). The impact of artificial intelligence on the professional field of public relations/communications management: Recent developments and opportunities, in artificial intelligence in public relations and communications: Cases, reflections, and predictions. Berlin: Quadriga University of Applied Sciences.

Tajfel, H., and Turner, J. C. (1979). "An integrative theory of intergroup conflict" in The social psychology of intergroup relations. eds. W. G. Austin and S. Worchel (Brooks/Cole: Monterey, CA), 33–47.

Thaler, R. H., and Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press.

The Toronto Declaration. Protecting the rights to equality and non-discrimination in machine learning systems. (2018). Available at: https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Zerfass, A., Hagelstein, J., and Tench, R. (2020). Artificial intelligence in communication management: a cross-national study on adoption and knowledge, impact, challenges and risks. *J. Commun. Manag.* 24, 377–389. doi: 10.1108/JCOM-10-2019-0137

frontiers | Frontiers in Artificial Intelligence

TYPE Policy and Practice Reviews
PUBLISHED 31 July 2025
DOI 10.3389/frai.2025.1569115

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
J. D. Opdyke,
DataMineit, LLC, United States
Hugh Lawson-Tancred,
Birkbeck University of London,
United Kingdom

*CORRESPONDENCE
Alexander Romanishyn
✉ a.romanishyn@ise-group.org

RECEIVED 31 January 2025
ACCEPTED 30 June 2025
PUBLISHED 31 July 2025

CITATION
Romanishyn A, Malytska O and Goncharuk V (2025) AI-driven disinformation: policy recommendations for democratic resilience.
Front. Artif. Intell. 8:1569115.
doi: 10.3389/frai.2025.1569115

AI-driven disinformation: policy recommendations for democratic resilience

Alexander Romanishyn*, Olena Malytska and Vitaliy Goncharuk

Think Tank ISE Group, Kyiv, Ukraine

The increasing integration of artificial intelligence (AI) into digital communication platforms has significantly transformed the landscape of information dissemination. Recent evidence indicates that AI-enabled tools, particularly generative models and engagement-optimization algorithms, play a central role in the production and amplification of disinformation. This phenomenon poses a direct challenge to democratic processes, as algorithmically amplified falsehoods systematically distort political information environments, erode public trust in institutions, and foster polarization – conditions that degrade democratic decision-making. The regulatory asymmetry between traditional media – historically subject to public oversight – and digital platforms exacerbates these vulnerabilities. This policy and practice review has three primary aims: (1) to document and analyze the role of AI in recent disinformation campaigns, (2) to assess the effectiveness and limitations of existing AI governance frameworks in mitigating disinformation risks, and (3) to formulate evidence-informed policy recommendations to strengthen institutional resilience. Drawing on qualitative analysis of case studies and regulatory trends, we argue for the urgent need to embed AI-specific oversight mechanisms within democratic governance systems. We recommend a multi-stakeholder approach involving platform accountability, enforceable regulatory harmonization across jurisdictions, and sustained civic education to foster digital literacy and cognitive resilience as defenses against malign information. Without such interventions, democratic processes risk becoming increasingly susceptible to manipulation, delegitimization, and systemic erosion.

KEYWORDS

AI, disinformation, deepfake, policy recommendation, AI regulation

1 Introduction

1.1 Empirical insights into disinformation dynamics

A growing body of empirical research has illuminated how disinformation proliferates across digital ecosystems, driven by a complex interplay of user psychology, algorithmic design, media structures, and emerging technologies. Rather than stemming from a single source or mechanism, disinformation spreads through multiple, reinforcing channels – each of which has been rigorously studied in recent years.

Vosoughi et al. (2018) conducted a landmark analysis of ~126,000 Twitter cascades and found that false news diffused "significantly farther, faster, deeper, and more broadly" than true news across all topics. False stories were typically more novel and emotionally evocative (e.g., inducing surprise or disgust), which likely made people more inclined to share them. Notably, this disparity was driven by human behavior rather than bots: the authors observed that automated accounts amplified true and false news at similar rates, implying that *human* users are more prone to spread falsehoods. This early large-scale evidence revealed a

fundamental vulnerability in the online information ecosystem – sensational misinformation appeals to users and propagates rapidly, posing a challenge for truth to keep up.

While humans are the primary propagators of viral falsehoods, subsequent research showed that automated "bot" accounts still play a significant amplifying role. Shao et al. (2018) analyzed 14 million tweets sharing hundreds of thousands of low-credibility news articles around the 2016 U. S. election and found that social bots had a *disproportionate* impact on spreading misinformation. In the crucial early moments of an article's life cycle, bots would aggressively spread links from false or low-quality sources – even targeting influential users via replies and mentions – to jumpstart viral momentum. Humans often then unwittingly reshared this content introduced by bots. Shao et al. conclude that curbing orchestrated bot networks could help dampen the initial wildfire-like spread of false stories online. In short, platform manipulation by bots was shown to *boost* the reach of fake news, even if the ultimate decisions to share lay with human users.

On Facebook, large-scale data suggests misinformation sharing is highly skewed to certain segments of the population. Guess et al. (2019) tracked people's Facebook activity during the 2016 election and found that only a small fraction of users accounted for the majority of fake news shares. Importantly, those most likely to share false news stories were disproportionately older adults: Americans over 65 shared several times more fake news articles on average than younger age groups, even after controlling for ideology and other factors. This robust age effect was not simply because older people use Facebook more or are more conservative – it persisted independent of partisanship. The findings point to specific demographic vulnerabilities in the spread of online misinformation. They suggest that digital media literacy gaps (particularly among seniors who did not grow up with the internet) could be a driving factor, and that interventions might be needed to support those populations. In sum, the propagation of fake news on social platforms is not uniform; it concentrates among certain user groups, which has implications for targeted counter-misinformation strategies.

Beyond user behavior and platform mechanics, the integrity of the information *supply* itself can contribute to widespread misinformation. Focusing on the COVID-19 pandemic, Parker et al. (2021) interviewed scientists and health communicators to identify systemic drivers of COVID misinformation in scientific communication. They found broad concern that certain failures in the scientific and publishing system – such as the publication of low-quality or biased research, limited public access to high-quality findings, and insufficient public understanding of science – greatly facilitated the spread of false or misleading health claims. For example, if preliminary or flawed studies are hyped without rigorous peer review, they can seed misinformation in public discourse. Parker et al. note that participants advocated a range of structural solutions: strengthening research standards and peer review, incentivizing careful science communication (e.g., translating findings for lay audiences), expanding open-access to credible research, and leveraging new technologies to better inform the public. There was even debate over the role of preprints – whether they exacerbate misinformation or help by rapidly disseminating data. The study's conclusions underscore that systemic failings in how science is produced and communicated can fuel misinformation crises. Thus, bolstering the transparency, rigor, and accessibility of scientific information is seen as a necessary policy response to prevent future infodemics. This complements the platform-focused studies by showing that the misinformation problem also roots in the upstream information ecosystem.

Encouragingly, empirical research has started to identify practical interventions to reduce the spread of disinformation/misinformation. Pennycook et al. (2021) demonstrated through large-scale experiments that simple behavioral "nudges" can significantly improve the quality of content people share online. They found that when social media users are subtly prompted to consider the accuracy of a news headline before deciding to share it, their likelihood of sharing false or misleading headlines drops substantially. In other words, many people do not *intend* to spread misinformation; rather, they often fail to think critically about truthfulness in the rush of online sharing. Pennycook et al.'s participants overwhelmingly said that sharing only accurate news is important to them, and the intervention helped align their sharing behavior with that value. This approach – shifting users' attention toward accuracy at key moments – proved effective across the political spectrum and did not rely on any partisan framing. The broader implication is that social platforms could implement low-cost, scalable accuracy checks or prompts to nudge users toward more mindful sharing. Such measures, informed by behavioral science, offer a promising complement to algorithmic or fact-checking-based solutions. They support the case for industry and policy initiatives that embed "friction" or reflection into the sharing process as a way to stem the tide of misinformation.

Finally, the advent of advanced AI technologies is transforming the misinformation landscape, raising new concerns that justify proactive policy intervention. A growing body of interdisciplinary research and case evidence provides a comprehensive review of emerging AI-driven disinformation techniques – from deepfake media to AI chatbots – and warns of their potential to dramatically amplify false narratives. Notably, today's state-of-the-art generative models (large language models) can produce politically relevant false news content that humans often cannot distinguish from real news. In one evaluation, some AI-generated election disinformation was indistinguishable from authentic human-written journalism in over half of instances. Moreover, AI systems can now mimic real individuals with uncanny realism; for example, language models have been shown to imitate the style of politicians or public figures with *greater perceived authenticity* than the persons' actual statements. Similar progress in deepfake video and audio means that one can forge convincing videos or voice recordings of public figures at scale. These developments lower the barrier for malicious actors to generate and disseminate false content on a massive scale, and they blur the boundaries between truth and fabrication in ways that could easily deceive the public. The rise of AI-generated disinformation has made the problem more acute than ever, necessitating robust countermeasures - from better detection technologies (e.g., deepfake detection, content provenance) to updated regulations on AI use - to safeguard information integrity. In short, the evolving capabilities of AI demand an equally evolving policy response.

Collectively, these empirical studies reinforce the urgent need for intervention at multiple levels – platform governance, user behavior, information integrity, and algorithmic transparency – to effectively counter AI-driven disinformation. The findings not only validate the systemic nature of the problem but also highlight concrete mechanisms through which disinformation proliferates and can potentially be mitigated. By grounding our analysis in this

growing body of research, we set the stage for a more granular examination of how specific disinformation modalities – such as deepfakes, bots, and synthetic personas – operate across different scales. The next section builds on these insights by categorizing disinformation risks and mapping their observable dynamics against global benchmarks to inform more targeted policy responses.

1.2 Mapping disinformation risk by category and scale

The rapid expansion of global internet and social media usage has substantially increased the surface area vulnerable to AI-driven disinformation campaigns. As of April 2025, an estimated 5.64 billion individuals – approximately 68.7% of the world's 8.21 billion population – were active internet users (DataReportal, 2025a, 2025b). Simultaneously, 5.31 billion social media accounts were in use, representing 64.7% of the global population (DataReportal, 2025a, 2025b). Average daily engagement remains high: users spent between 143 and 147 min per day on social media platforms during early 2025 (SOAX, 2025; Exploding Topics, 2025; BusinessDasher, 2024).

This level of global digital saturation offers a fertile environment for disinformation to propagate rapidly, especially as generative AI systems enable low-cost, scalable content production and targeting. Against this backdrop, it becomes imperative to benchmark observed disinformation trends against these broader usage patterns. The sections below present category-specific trends – ranging from deepfakes and bots to synthetic identities and real-time disinformation – contextualized within global benchmarks to assess scope, scale, and systemic risk (see Table 1).

Broadly, the benchmark comparisons confirm that observed trends are part of a larger, accelerating wave of AI-fueled disinformation. In many cases, the specific increases noted (e.g., a fivefold rise in deepfakes, or tens of millions of fake identities) mirror global surges in those phenomena (NewsGuard, 2024; Sumsub, 2024). This alignment strengthens our confidence that the trends are real and significant – not just isolated anomalies – and underscores the urgency for interventions.

Several notable insights emerge from examining the metrics in light of benchmarks:

The deepfakes category illustrates an arms race between AI capabilities and detection. A 5 × local increase in deepfake content is contextualized by exponential global growth (550% since 2019) in known deepfake videos (Security.org, 2024; MarketsandMarkets, 2024). Crucially, deepfakes are transitioning from niche (mostly pornographic content) to mainstream weaponization in scams, politics, and malign influence. Meanwhile, the volume of deepfakes online is growing exponentially – around *half a million* deepfake videos were shared on social media in 2023, and projections show up to 8 million by 2025. This arms race between deepfake generators and detectors underscores the urgent need for countermeasures (Moore, 2024). The fact that over 6% of fraud incidents now involve deepfakes (ACI Worldwide, 2024) signals to policymakers that this is no longer a theoretical threat – it is actively undermining financial and information integrity. Implication: There is a pressing need for better deepfake detection tools and regulatory frameworks. Policy options include mandating provenance watermarks for AI-generated media and criminalizing malicious deepfake use (European Parliament, 2022). Investing in R&D for detection (e.g., deepfake forensics, authenticated media pipelines) is critical. The benchmarks also suggest international cooperation is needed: deepfake fraud and disinformation are rising across all regions (e.g., +1740% in N. America, +780% in Europe in one year) (Redline Digital, 2023), so solutions must be global in scope.

1.2.1 Generative text oversupply

The text generation trends show AI-written disinformation now spreads widely and often evades traditional detection. Our benchmarks confirm that AI-driven "fake news" sites grew tenfold in one year (NewsGuard, 2024), flooding the infosphere with low-cost, algorithmically generated propaganda. This mass production of content, combined with the known fact that falsehoods spread faster than truths on social networks (Vosoughi et al., 2018), creates a perfect storm: even moderately convincing AI fakes can achieve wide circulation before fact-checkers can respond. Implication: Traditional moderation (reactively deleting false posts) may be insufficient – we need proactive and AI-assisted countermeasures. Possibilities include AI content provenance checks, real-time content authenticity scoring, and "counter-LLMs" deployed to detect AI-generated text patterns. Policies encouraging transparency (such as requiring labeling of AI-generated political ads or narratives) could help. The stark benchmark that NewsGuard already counted 1,200 + AI content websites by 2024 also raises the issue of media literacy: users must be educated to scrutinize sources, as many websites may now be effectively "content farms" run by bots or generative models. The lack of inferential controls in our data is mitigated by these comparisons – it is clear that the rise in AI text disinformation we observed is part of a systemic transformation of the information landscape.

1.2.2 Scale of fake personas

It's noted that millions of Synthetic Identities deployed in disinformation campaigns, and benchmarks reinforce how widespread this tactic has become. The Federal Reserve (Boston Fed, 2024) reports $35 billion in losses in 2023 from synthetic identity fraud (often related to financial crimes). While financial fraud is one aspect, the same fake persona generation techniques are being repurposed for information operations – fake activists, fake journalists, fake grassroots groups. The 5 × jump in digital identity forgeries in just two years underscores the impact of generative AI in automating the creation of credible-looking personas (complete with AI-generated profile photos and even "deepfake voices") (Sumsub, 2024). Implication: From a policy perspective, strengthening digital identity verification is key. Social media platforms and messaging services may need stricter "know-your-customer" rules for high-reach accounts or for political advertisers, to prevent large-scale bot armies with AI faces from multiplying. There is a trade-off with privacy/anonymity, but benchmarks show the pendulum has swung toward chaos – millions of fake identities can now be marshaled to distort the public discourse. Improving authentication (blue-check verification reforms, multi-factor checks for accounts) and tracking coordinated fake account networks (Meta, 2023) should be a priority. Additionally, collaborative efforts like sharing bot/fake account blacklists across platforms could reduce cross-platform abuse.

TABLE 1 Trends in context of global digital dynamics (2019–2025).

Trend category	Observed growth	Benchmark insight	Relevance & interpretation
Deepfakes	5 × increase in deepfake content; +2,137% fraud attempts in North America	550% growth in global deepfake videos (2019–2023); 6.5% of all fraud in 2023 involved deepfakes	Mirrors an exponential content surge. Fraudulent deepfakes have shifted from niche uses (e.g., non-consensual porn) to mainstream scams and political impersonations, fueling nearly 40% of high-value fraud cases in 2024. Highlights a growing cross-sector risk, calling for stronger detection and legal safeguards.
Text generation	AI-generated news outpaced detection; fooled >50% of human evaluators	10 × increase in AI-generated fake news sites; 1,200 + sites by 2024; false news 70% more likely to spread than true	Illustrates rapid scale-up in misinformation production. Virality benchmarks show systemic threat to information integrity.
Synthetic identities	12.8 M + fake personas used for influence ops	$35B in synthetic ID fraud losses; 5 × rise in forged identities; 1.3B + fake accounts disabled quarterly by Meta	Highlights ease of creating believable AI personas and difficulty in platform-level filtering. Economic cost shows systemic vulnerability.
Bots (automation)	~25% of Twitter activity was bot-driven	50% of total web traffic by bots (2023); ~30–37% from "bad bots"	Confirms that bot-driven amplification is widespread and not platform-specific. Confounds detection, skews public discourse.
Microtargeting	AI-enhanced propaganda reached ~34% of users	75% of firms use AI targeting; 3 × higher conversion for personalized ads; $366B digital ad market	Personalized messaging has become default. Political targeting mimics commercial tactics – raises risks of manipulation and democratic erosion.
Real-time disinformation	Fact-check latency shrank from 45 to 15 min	Viral false news spreads ~6 × faster than truth; initial exposure shapes long-term belief	Even 15 min is enough to embed false narratives. Detection speed must be paired with pre-bunking and real-time mitigation tools.

1.2.3 Pervasive bot automation

In the Bots category, the estimate (~25% of social media interactions in some domains are driven by bots) is backed by global data showing nearly half of all web traffic is non-human (Imperva/Statista, 2024). This indicates that what we observe on one platform (e.g., a quarter of Twitter content being bot-driven) is symptomatic of a wider phenomenon. Bots – especially "bad bots" – are now deeply entrenched in social media ecosystems, often orchestrated to amplify disinformation (Zhang et al., 2023). For instance, during health crises and elections, researchers have found bot networks boosting polarizing or false narratives (Shao et al., 2018). Implication: The prevalence of bots calls for strengthening platform defenses and perhaps regulatory oversight on transparency. Platforms should be encouraged (or required) to detect and label bot accounts (e.g., through robust use of tools like botometer or in-house AI systems) and to shut down coordinated inauthentic networks proactively. Benchmarks show even verified channels can be co-opted by bots (e.g., the mention of "verified bot" problems on Twitter/X), so verification processes need updates to stay ahead of AI-fueled fakery. From a policy angle, one idea is introducing "bot disclosure" laws – requiring automated accounts to self-identify as such – which some jurisdictions have explored. Additionally, research must continue on distinguishing AI-driven behavior from human behavior online; the battle is dynamic (as bots get more human-like with AI, detection must get more sophisticated).

1.2.4 Microtargeting and personalized propaganda

The Microtargeting trends highlight that AI can tailor messages to exploit individuals' psychological profiles at scale. The observed stat – targeted campaigns influencing 34% of users – is plausible given that most advertisers already use AI-driven microtargeting and see greatly improved engagement. Benchmarks on ad performance (e.g., targeted ads yielding 10 × higher click-through with retargeting) suggest why malactors also microtarget: personalized disinformation is simply more effective at persuading or manipulating than one-size-fits-all propaganda (AI Marketing Advances, 2023; Redline Digital, 2023). Implication: There are critical policy and ethical questions here. Should microtargeted political ads be limited or banned? (For example, the EU has considered restrictions on microtargeting voters with political messages.) At minimum, transparency measures are needed – users should know why they are seeing certain political or issue-based content, and researchers should have access to platform data on ad targeting (as enabled by ad libraries in the EU's Digital Services Act). Moreover, digital literacy programs should teach citizens that the news or ads they see might be algorithmically curated specifically for them, which could help inoculate against blindly trusting tailored misinformation. Technical defenses might involve detecting when malicious actors use advertising tools or algorithmic boosting for propaganda – for instance, unusual patterns in ad buys or content that is disproportionately targeted at vulnerable groups could trigger audits. The benchmarks reinforce that microtargeting is not inherently nefarious (it's now a standard marketing practice), but when used to propagate false or extremist content, its high success rate becomes a threat to the democratic discourse.

1.2.5 Speed of disinformation vs. response

The Real-Time Disinformation metrics indicate some progress – response times to viral falsities have improved (15 min on average in 2023, whereas in 2021 it often took 30–45 min or more for corrections to emerge) (Disinformation Research, 2023). However, when juxtaposed with the benchmark that false news can achieve massive spread in minutes, it's evident that even a 15-min lag is problematic. Malicious actors capitalize on breaking news moments – for example, immediately after a disaster or major announcement – to inject falsehoods that ride the momentum before facts are verified. Our case of adversaries "weaponizing speed" is exemplified by real incidents (e.g., a fake news report causing brief market chaos before being

debunked) (Hunter J., 2024). Implication: Real-time detection and intervention must be a centerpiece of counter-disinformation strategy. This could involve algorithms monitoring spikes in certain keywords or sentiment shifts that often accompany fake virality, and then alerting moderators or posting contextual warnings in nearly real-time. Collaboration between platforms and independent fact-checkers/press agencies is also key – e.g. WhatsApp's pilot projects with rapid rumor quashing, or Google working with health authorities to counter misinformation spurts. The 2023 landscape saw the emergence of community-driven efforts (like Community Notes on X/Twitter) that can sometimes provide corrective context within hours of a misleading post – a positive development, but still not fast enough in many cases. Going forward, "pre-bunking" strategies (releasing forewarnings about likely false narratives before they spread) and crisis-response communication protocols (so that official sources can flood the zone with accurate info quickly during emergencies) are needed. The improvements from 45 → 15 min show it's possible to shorten reaction time; benchmarks like the MIT study (false news 6 × faster) show we must get even faster (Vosoughi et al., 2018). Policymakers might support the creation of real-time misinformation monitoring centers, possibly run by coalitions of tech companies, civil society, and governments, to coordinate swift responses to emerging disinformation campaigns.

In summary, benchmarking the observed disinformation trends against global data confirms that these patterns are not isolated incidents but part of a widespread and accelerating phenomenon. While the present analysis does not employ formal inferential controls, the convergence with international benchmarks offers external validation of the identified trends. For instance, a fivefold local increase in deepfake content closely parallels a tenfold global rise, reinforcing the urgency of addressing this threat. Likewise, the proliferation of AI-generated news sites and synthetic identities highlights the systemic nature of disinformation vectors across technological, social, and political domains. These benchmark comparisons provide essential interpretive context, clarifying which categories represent the highest systemic risk and guiding where targeted interventions may be most effective.

Building on these findings, the following section critically examines the existing policy landscape and mitigation strategies. It assesses the strengths and limitations of current regulatory approaches, identifies persistent gaps, and outlines evidence-based recommendations to enhance societal and institutional resilience. Given the transnational character of AI-enabled disinformation, the analysis emphasizes the need for coordinated, forward-looking governance mechanisms that are both adaptive and inclusive.

2 Sections on assessment of policy/guidelines options and implications

2.1 Global landscape of AI regulations and country specific cases

The development and implementation of artificial intelligence (AI) regulations have gained significant momentum worldwide. Since 2017, 69 countries have collectively adopted over 800 AI regulations (World Health Organization, 2018). Regulations aim to address issues such as bias, discrimination, privacy, and security. Most regulatory frameworks seek to foster AI development while safeguarding individual rights and societal interests as well as fighting disinformation. Some regulations, like the EU AI Act, will apply to AI systems used within their jurisdiction, regardless of the provider's location (European Parliament, 2022).

2.1.1 European Union—the AI act (2024)

The EU's artificial intelligence act is a comprehensive risk-based framework that categorizes AI applications by risk level. It bans certain "unacceptable risk" uses of AI (for example, social scoring and real-time biometric surveillance in public), imposes strict requirements on "high-risk" systems (such as those used in critical infrastructure or law enforcement), and places minimal restrictions on low-risk applications. Importantly, the AI Act includes transparency mandates relevant to disinformation: providers of generative AI must disclose AI-generated content and ensure compliance with EU copyright laws. High-impact general-purpose AI models will face additional evaluations. The Act was adopted in 2024, with most provisions enforceable starting in 2026, and is supported by significant EU funding (around €1 billion per year) to encourage compliant AI development. Notably, the regulation has extraterritorial reach, applying to any AI system used in the EU market regardless of the provider's origin.

2.1.2 United States—sectoral and state-level approach

The United States lacks a single comprehensive AI law at the federal level. Instead, it relies on a patchwork of initiatives and guidelines. The National AI Initiative Act of 2021 established a coordinated strategy for federal investment in AI research and development, and the National Institute of Standards and Technology (NIST) released an AI Risk Management Framework to promote trustworthy AI practices. However, there is no dedicated federal law addressing AI in areas like privacy or disinformation. Enforcement is carried out through sector-specific regulations and state laws – for example, New York City's Local Law 144 (effective July 2023) regulates the use of AI in hiring decisions, and the Federal Trade Commission has warned it will police deceptive commercial uses of AI under its broad consumer protection mandate. This decentralized approach has led to gaps and inconsistencies. The absence of a federal AI-specific privacy or content law means digital platforms in the U.S. primarily govern AI-driven disinformation through their own policies, under general oversight such as anti-fraud and election laws. The result is a less uniform defense against AI misuse, as rules can differ significantly by state and sector.

2.1.3 Canada – the artificial intelligence and data act (proposed 2022)

Canada has pursued a balanced approach with its proposed Artificial Intelligence and Data Act (AIDA), introduced in June 2022 as part of a broader Digital Charter Implementation Act. AIDA would establish common requirements for AI systems, especially high-impact applications, focusing on principles like transparency, accountability, and human oversight. As of late 2024, AIDA has not yet been passed into law. In the meantime, Canada has invested heavily in AI innovation (over $1 billion by 2020) and released guidelines such as the Directive on Automated Decision-Making for government use of AI. The intent is to encourage AI advancement

(Canada is home to a robust AI research community) while guarding against harms like biased or unsafe AI outcomes. If enacted, AIDA is expected to introduce mandatory assessments and monitoring for risky AI systems, which could encompass tools that spread or detect disinformation.

2.1.4 United Kingdom – National AI Strategy (2021)

The UK's approach, outlined in its National AI Strategy, is characterized by a "pro-innovation" stance that so far avoids sweeping new AI-specific legislation. Instead, the UK has articulated high-level principles such as safety, fairness, and accountability to guide AI development. It empowers existing regulators (in finance, healthcare, etc.) to tailor AI guidance for their sectors rather than creating a single AI regulator. Starting in 2023 and into 2024, the UK government has been evaluating how to implement these principles, including whether to introduce light-touch regulations in specific areas. For instance, Britain has considered regulations on deepfakes and online harms as part of its Online Safety Bill, and it hosted a global AI Safety Summit in late 2024 to discuss international coordination. However, as of early 2025, the UK relies on general data protection law (the UK GDPR), consumer protection law, and voluntary industry measures to handle AI-driven disinformation. This decentralized approach gives industry more flexibility, but some critics worry it may lag behind the curve on fast-moving threats like synthetic media manipulation.

2.1.5 India – IT rules and content takedown (2021–2023)

In India, policymakers have approached disinformation largely through amendments to information technology regulations. The Information Technology Rules, 2021 (under the IT Act, 2000) mandate social media platforms to remove false or misleading content flagged as such by the government. The Intermediary Guidelines and Digital Media Ethics Code established under these rules obligates tech companies to actively curb the spread of misinformation on their services. In 2023, India introduced further measures by amending the IT Rules to create a government-run fact-checking unit tasked with identifying false information related to government policies. As a result, platforms like Twitter, Facebook, and WhatsApp are now required to comply with official takedown requests targeting content deemed "fake news" by authorities (Mehrotra and Upadhyay, 2025). These legal mechanisms showcase India's aggressive stance in forcing platform compliance; however, they have raised concerns among free speech advocates. Granting the government broad power to determine truth can risk overreach and censorship, potentially stifling legitimate dissent (Access Now, 2023). India's experience highlights the tension in regulating disinformation in a democracy – how to curb dangerous falsehoods at scale without undermining civil liberties. The high volume of content in India (over 800 million internet users by 2025) also makes enforcement difficult, and misinformation (often via WhatsApp forwards in myriad local languages) has continued to spark violence and social unrest in recent years. Thus, while India's regulatory strategy enables swift removal of content, it also underscores the need for complementary approaches like community fact-checking and media literacy to address root causes of belief in rumors.

2.1.6 China – deep synthesis regulations (2023)

China's authoritarian information control regime has taken a markedly different approach, focusing on stringent preemptive regulations to prevent AI misuse. The Cyberspace Administration of China (CAC) implemented pioneering rules on "deep synthesis" technology that took effect in January 2023 – one of the world's first comprehensive laws targeting deepfakes. These regulations prohibit the use of deepfake or other generative AI technology to produce and disseminate "fake news" or content that could disrupt economic or social order. Service providers offering generative AI tools are required to authenticate users' real identities and to ensure their algorithms do not generate prohibited content. Critically, the rules mandate that any AI-generated or AI-manipulated content must be clearly labeled as such, both through visible markers in the content (e.g., watermarks or disclaimers) and through embedded metadata (China Daily, 2025). This dual labeling requirement is intended to ensure public awareness that a piece of media is synthetic and to facilitate traceability via hidden "fingerprints" in case the content is reposted or altered. Chinese platforms and internet services are held liable for enforcing these rules – meaning they must detect and remove unlabeled deepfakes and report violators to authorities. The Chinese government's approach emphasizes a security-first, censorship-heavy model: it leverages its centralized control over the internet to preemptively block and punish the creation of AI-driven disinformation within its borders. At the same time, China has been known to deploy disinformation abroad via state media and covert influence operations. Thus, domestically China shows one extreme of a regulatory regime (mandatory labeling and strict prohibition), though such an approach is enabled by its broader curbs on speech and would likely not be palatable in open societies.

2.1.7 Brazil (and South America) – proposed "Fake News" law (2020–2023)

Across South America, governments have also grappled with how to stem harmful disinformation, which in countries like Brazil spreads rapidly through platforms like WhatsApp and YouTube. Brazil's experience is illustrative. In 2020, lawmakers introduced Bill No. 2630, dubbed the "Fake News Bill," to address online misinformation and abuse. After delays, the bill regained momentum in 2023 with significant revisions. The far-reaching proposal would overhaul internet liability rules by establishing new "duty of care" obligations for tech platforms regarding content moderation (Freedom House, 2023). Rather than waiting for court orders, platforms would be required to actively monitor, find, and remove illegal or harmful content (including disinformation) or face hefty fines (Paul, 2023). The bill also includes provisions to curb inauthentic behavior, such as requiring phone number registration for social media accounts (to limit anonymous bot networks) and increasing transparency of political advertising. Major tech companies have fiercely opposed the legislation, arguing that some provisions threaten encryption and free expression (for instance, by potentially requiring WhatsApp to trace forwarded messages, which conflicts with end-to-end encryption). Under pressure, a vote on PL 2630 was postponed in 2023, and as of early 2025 it remains pending in Brazil's Congress (Freedom House, 2023). Nonetheless, Brazil's judiciary stepped into the breach during recent elections – the Electoral Court (TSE) ordered the removal of thousands of pieces of false content and even temporarily suspended messaging apps that failed to control rampant misinformation. Other

South American countries have taken smaller steps: Argentina launched media literacy programs; Colombia and Peru have set up anti-fake-news observatories. The trajectory in the region suggests a trend toward imposing greater accountability on platforms for content spread, while trying to balance the demands of democracy and free speech. Brazil's proposed law, in particular, if passed, would be among the world's strongest social media regulations against disinformation, potentially influencing other democracies facing similar challenges.

Overall, the global regulatory landscape for AI and disinformation is fragmented. The EU's comprehensive and stringent regime contrasts with the more *ad hoc* or principles-based approaches in the US and UK, as well as the heavy censorship model in China. Such divergence can complicate efforts to combat AI misuse internationally. Gaps between jurisdictions create opportunities for malicious actors to engage in regulatory arbitrage – exploiting the most lenient environment to base their operations or technological development. For example, a disinformation website flagged in Europe can move its hosting to a country with weaker rules; AI developers can open-source a potentially harmful model from a location with minimal oversight. Similarly, inconsistent standards (like on political deepfake ads) mean content banned in one country may still reach audiences elsewhere. This patchwork of rules underscores a need for greater international coordination, which we address later in our recommendations and future directions.

2.2 AI regulations and their implications for disinformation

Despite new regulations, significant challenges remain in addressing AI-fueled disinformation. One major issue is global fragmentation of AI governance. The lack of uniform rules across jurisdictions creates opportunities for malicious actors to engage in regulatory arbitrage – exploiting gaps by operating from countries with lax or no AI oversight. For example, an organization banned from deploying certain AI-generated content in the EU could base its operations in a country with weaker laws and still target EU audiences online. Inconsistent standards also mean that what one nation labels and detects as AI-generated disinformation may go unrecognized elsewhere. Companies attempting to police disinformation globally face a complex compliance puzzle, needing to navigate multiple legal regimes simultaneously. This patchwork of regulations can slow down cross-border responses to online campaigns and limit collaboration. International efforts to counter disinformation – such as sharing threat intelligence or coordinated content takedowns – are more difficult without a harmonized legal foundation.

Enforcement also looms large as a concern. Laws on paper do not automatically translate to effective action. Resource constraints, jurisdictional limits, and Big Tech's resistance can all undermine enforcement. Many countries' regulatory agencies lack the technical expertise or manpower to audit complex AI systems or to continuously monitor platforms for compliance. In the case of India's aggressive takedown rules, enforcement relies on platforms' willingness and ability to quickly remove flagged content; encrypted messaging services pose a further obstacle. In open societies, enforcement must also respect free speech and avoid political abuse – a difficult balance when governments themselves can become arbiters of truth. There is an inherent tension between controlling disinformation and upholding democratic freedoms. Overly heavy-handed laws risk censorship and could drive misinformation to harder-to-monitor channels (like private groups or decentralized networks).

Meanwhile, a technological arms race is underway between those creating disinformation and those attempting to counter it. Tighter regulations, while intended to curb abuse, can inadvertently fuel this race. As platforms and regulators impose new constraints, disinformation actors often respond by developing more advanced and evasive AI tools. For instance, mandated content labeling may incentivize adversaries to refine generative models that produce outputs indistinguishable from authentic content. In turn, governments and private actors race to develop more sophisticated defensive technologies to detect and neutralize these threats. This dynamic escalates the complexity of both offense and defense: bots trained to mimic human behavior can bypass AI-based filters, and hyper-realistic deepfakes may outpace current detection systems.

The rapid evolution of generative AI compounds this challenge, frequently outpacing legislative and regulatory cycles. By the time a governance framework is enacted, the threat landscape may have already shifted. Effective policy responses must therefore move beyond reactive regulation or blanket investment and instead prioritize agile, pre-competitive collaboration between research institutions, civil society, and industry. Such frameworks could include shared threat modeling environments, sandboxed co-development of detection tools, and adaptive policy toolkits designed to evolve in parallel with AI capabilities.

Efforts to rein in disinformation must also consider freedom of speech. Measures like automated content filters or legal penalties for spreading "fake news" risk encroaching on legitimate expression if not carefully calibrated. AI-driven content moderation, for instance, can mistakenly flag satire, opinion, or contextually complex posts as disinformation. Overzealous or poorly tuned algorithms might over-censor and suppress valid discussions, raising concerns about violating free speech rights. The lack of human judgment in some AI moderation tools means nuance can be lost – what is misleading in one context might be valid in another, and current AI may struggle with such distinctions. Additionally, if governments mandate certain AI filters, there is a danger that those systems could be biased towards particular political or cultural viewpoints, intentionally or not, thereby silencing dissenting voices under the guise of combating falsity. Maintaining transparency and avenues for appeal in content moderation decisions is thus critical to uphold democratic values even as we try to clean up the information space.

Another set of issues revolves around data protection and privacy. Paradoxically, laws like the European GDPR – which protect users' personal data and grant rights over automated profiling – can impede the fight against disinformation. Effective AI detection of fake accounts or targeted misinformation often requires analyzing large amounts of user data (to spot inauthentic behavior patterns or trace how false stories spread through networks). Privacy regulations restrict access to some of this data or require anonymization that might reduce its utility. For example, an AI system might be less effective at identifying a network of coordinated fake profiles if it cannot easily aggregate personal metadata due to legal constraints. GDPR also gives users the right to opt out of automated decision-making or to demand explanations for it; platforms, fearing liability, might limit their use of AI moderation tools or throttle back automation to avoid infringing those rights. This could constrain

content moderation efforts just when AI's speed and scale would be most useful. Policymakers thus face a difficult task in reconciling robust privacy protections with agile anti-disinformation mechanisms.

In addition, there are practical and resource challenges. Disinformation is a cross-border, cross-platform problem, but enforcement mechanisms are typically national. Even when countries agree on principles, coordinating actions (such as shutting down a global botnet or sanctioning a foreign propaganda outlet) is cumbersome. Companies that operate internationally must deal with not only multiple laws but also sometimes conflicting demands (for instance, one country may demand certain content be removed as "fake," while another country's law protects that content). Cross-border cooperation among regulators is still nascent, and mechanisms for rapid information sharing or joint action are limited. Furthermore, not all organizations have the capacity to implement the latest AI defenses. Large social media firms invest heavily in AI moderation and teams of experts, but smaller platforms or local media outlets often lack such resources. This creates a weak link that disinformants can exploit, by spreading falsehoods on less moderated services and then pushing that content into mainstream discussion. Finally, AI tools themselves can carry biases that affect what is labeled disinformation – if a detection algorithm is trained on skewed data, it might overlook certain languages or communities, thereby missing targeted misinformation campaigns in those segments. All these additional challenges highlight that combating AI-driven disinformation requires not just laws and algorithms, but also capacity building, international norms, and continual refinement of both technological and policy approaches.

In summary, current AI regulatory efforts, while a crucial start, have notable limitations when confronting disinformation. Fragmented governance allows malicious actors to maneuver around restrictions, and even within regulated spaces, transparency, enforcement, and rights-balancing pose difficulties. A coordinated, adaptive strategy is needed – one that harmonizes laws across borders, updates rules in step with technological advances, safeguards fundamental freedoms, and supports both major and minor stakeholders in the information ecosystem. The next section will propose actionable policy recommendations that build on this analysis, aiming to strengthen our collective ability to counter AI-powered falsehoods.

3 Actionable recommendations

3.1 Limitation in AI regulation vs. policy recommendations

As we have explored the global landscape of AI regulations, country-specific approaches, and the implications of AI regulations on disinformation, it becomes evident that there are significant limitations in current AI regulatory frameworks. These limitations necessitate a more nuanced and forward-thinking approach to policy recommendations. The complex interplay between rapidly advancing AI technologies, diverse national interests, and the ever-evolving nature of disinformation presents unique challenges that cannot be adequately addressed by existing regulatory measures alone.

Moreover, the challenges in enforcing transparency, the potential for a technological arms race in disinformation, and the delicate balance between regulation and innovation underscore the limitations of current regulatory efforts.

In light of these challenges, it is crucial to examine the limitations of existing AI regulations and propose policy recommendations that can effectively address the multifaceted issues surrounding AI governance, particularly in the context of combating disinformation. The following section will delve into these limitations and present a set of policy recommendations designed to bridge the gaps in current regulatory frameworks, foster responsible AI development, and enhance our collective ability to mitigate the risks associated with AI-driven disinformation (see Table 2).

3.2 From limitations to a phased governance framework

While existing AI regulations have begun to address certain risks, particularly in content moderation and transparency, substantial gaps remain in managing the dynamic and adversarial nature of disinformation ecosystems. The preceding analysis underscored several structural limitations, including regulatory lag, jurisdictional fragmentation, and the challenge of operationalizing adaptability in policy instruments. To bridge these gaps, we propose a structured set of policy interventions calibrated to the velocity of AI advancement and the evolving threat landscape.

The following matrix presents a temporal-impact framework that categorizes the proposed policy recommendations by their expected timeframe for implementation (short-term, transitional, long-term) and level of systemic impact. This approach facilitates strategic sequencing, allowing policymakers, platforms, and civil society actors to prioritize actions that are both immediately actionable and scalable over time. It also provides a roadmap for balancing urgent mitigation needs with sustainable governance mechanisms (see Table 3).

The analysis of existing AI regulations and their limitations in addressing disinformation reveals several critical gaps that require high-impact policy recommendations. These recommendations are designed to address the shortcomings of current regulatory frameworks and provide a more robust approach to combating AI-driven disinformation.

3.2.1 Mandate bias audits and diverse datasets

To enhance the operational integrity of AI systems used in content moderation and detection – particularly in high-impact applications influencing public discourse – we recommend a risk-based framework for algorithmic auditing and data inclusivity verification. Rather than imposing uniform compliance burdens, regulatory efforts should prioritize functional performance across linguistic, demographic, and regional variables, thereby ensuring that disinformation targeting underrepresented groups is not systematically overlooked. A tiered approach to auditing – similar to cybersecurity red-teaming – should be encouraged through lightweight, third-party evaluations based on internationally recognized standards. Developers may be incentivized to engage in voluntary self-assessments, supported by transparency reporting and regulatory safe harbors.

This recommendation is not intended to expand regulatory complexity, but to address a critical performance gap: systems trained on narrow datasets are less capable of detecting non-dominant

TABLE 2 Policy limitations and targeted recommendations for AI-driven disinformation mitigation.

Limitation	Policy recommendation	Guidance for implementation
Global fragmentation	Promote international cooperation and harmonization of AI regulations.	- Establish a global AI governance body under the UN or OECD - Develop an international AI treaty or convention. - Create regional AI regulatory frameworks. - Implement cross-border information-sharing mechanisms.
Regulatory loopholes in lax jurisdictions	Develop incentives for nations to adopt robust AI governance standards.	- Create an international AI compliance rating system. - Offer technical assistance for developing nations. - Implement trade incentives for adhering to global standards. - Establish a global AI governance fund.
Compliance burdens for small businesses	Introduce tiered regulatory frameworks based on company size and risk profile.	- Simplify compliance procedures for SMEs. - Provide government-sponsored AI compliance toolkits and resources. - Provide tax incentives for governance investments.
Technological arms race with malicious actors	Facilitate agile, pre-competitive collaboration to develop adaptive defenses.	- Fund interdisciplinary research at the intersection of AI, behavioral science, and security. - Develop shared threat modeling environments across public-private-academic sectors. - Deploy modular, updatable AI-powered detection systems with open benchmarking. - Introduce policy toolkits that can evolve with technological developments through iterative stakeholder engagement.
Balancing free speech and regulation	Develop transparent content moderation guidelines that prioritize human rights.	- Establish multi-stakeholder councils to create content moderation standards. - Implement appeal mechanisms for content removal decisions. - Require transparency reports from AI-powered systems. - Develop AI literacy programs.
Dataset biases in AI systems	Mandate bias audits and diverse dataset requirements for AI training.	- Establish industry standards for AI bias testing and mitigation. - Create diverse, open-source dataset repositories. - Require companies to disclose demographic data in training datasets. - Conduct algorithmic impact assessments for high-risk systems.
Contextual understanding challenges in content moderation	Invest in explainable AI (XAI) systems to improve decision- making.	- Fund research into context-aware AI models. - Develop benchmarks for contextual AI understanding. - Use human-in-the-loop systems for complex decisions. - Create guidelines for AI-human collaboration in moderation.
Evolving nature of disinformation tactics	Establish adaptive regulatory mechanisms that evolve with technological advancements.	- Create an international AI threat monitoring system. - Regularly review and update AI regulations. - Conduct scenario planning for future threats. - Form cross-sector working groups to address emerging challenges. - Create regulatory sandboxes.
Polarization and public trust issues	Promote cognitive resilience digital literacy campaigns about disinformation risks.	- Integrate AI and digital literacy and cognitive resilience into school curricula. - Launch awareness campaigns on social media platforms. - Provide easy-to-use fact-checking tools. - Create community programs to combat misinformation.
Cross-border enforcement challenges	Develop international cooperation mechanisms for AI regulation enforcement.	- Establish bilateral and multilateral agreements. - Create international dispute resolution mechanisms. - Implement cross-border data-sharing protocols. - Standardize extraterritorial AI regulation procedures. - Establish a global AI security alliance focused on disinformation resilience.

language content and culturally specific manipulation tactics. A measured and outcome-focused auditing regime can improve detection precision while maintaining innovation incentives and public trust.

3.2.2 Develop international cooperation mechanisms

Given the global nature of online platforms and influence operations, international governance cooperation is essential to close the gaps that single-nation policies leave open. We recommend the establishment of a transnational AI oversight body or coalition – potentially under the auspices of the United Nations or a consortium like the OECD – that facilitates coordination among national regulators. Such a body could maintain an up-to-date repository of AI-driven disinformation threats and the techniques used to counter them, allowing countries to share best practices in real time. It should create standardized protocols for information-sharing, so that if one country discovers a network of deepfake accounts, others can be alerted swiftly through an agreed channel. Joint cyber exercises or simulations could be run to improve collective readiness for major

TABLE 3 Phased policy interventions for AI governance: Timeframe and impact matrix.

Timeframe	Moderate impact	High impact
Short-term (immediate action)	- Introduce tiered regulatory frameworks based on company size and risk profile. - Promote cognitive resilience and digital literacy campaigns about disinformation risks.	- Mandate bias audits and diverse dataset requirements for AI training - Develop international cooperation mechanisms for AI regulation enforcement.
From short to long (transitional)	- Invest in explainable AI (XAI) systems to improve decision-making. - Develop incentives for nations to adopt robust AI governance standards. - Promote international cooperation and harmonization of AI regulations.	- Establish adaptive regulatory mechanisms that evolve with technological advancements.
Long-term (strategic planning)	- Address cross-border enforcement challenges - Through global governance frameworks.	- Facilitate agile, pre-competitive collaboration to develop adaptive defenses.

disinformation attacks (for instance, ahead of international elections or referendums). Furthermore, aligning certification and compliance standards for AI systems across borders would reduce adversaries' ability to exploit weak links. We propose pursuing mutual recognition agreements for AI audit and certification regimes. In practice, this means if an AI model is certified as safe and transparent in one jurisdiction, others accept that certification – and conversely, models or services flagged as malicious in one place can be rapidly restricted elsewhere. To address regulatory loopholes in lax jurisdictions, the international coalition could also introduce incentive structures for nations to adopt robust AI governance. This might include an AI governance rating or index and tying development aid, trade benefits, or membership in certain international forums to improvements in AI regulatory standards. Technical assistance programs could help developing countries craft and enforce AI rules so they do not become unwitting safe havens for disinformation operations. Over time, these measures foster a more harmonized global approach, making it harder for disinformation agents to simply relocate their activities to avoid scrutiny.

To institutionalize these efforts, we advocate for the creation of a Global AI Security Alliance – a network of regulatory authorities, research institutions, and digital platform providers dedicated to proactive defense and regulatory convergence. This alliance should be initiated by a core group of digitally advanced democracies (e.g., EU, U.S., Japan, Australia) via a multilateral memorandum of understanding (MoU). Operational capacity would be structured around three interconnected pillars: (1) Shared Intelligence Infrastructure, for exchanging real-time data on emerging AI threats and synthetic content across jurisdictions; (2) Joint R&D Accelerator, to support cross-sector consortia building modular, adaptive detection systems and provenance verification tools; (3) Policy Harmonization Track, to align AI oversight standards through regulatory sandboxing, mutual recognition of audits, and best-practice exchange.

Anchoring this initiative in a multilateral framework (e.g., G7, OECD, or UNESCO) would enhance legitimacy, scalability, and interoperability. To address regulatory loopholes in lax jurisdictions, the alliance could introduce incentive structures – such as a global AI governance index or linking development aid, trade benefits, and digital access to regulatory compliance. Technical assistance programs should also be developed to support capacity-building in lower-resourced jurisdictions. These coordinated mechanisms would form a robust global governance architecture, making it increasingly difficult for disinformation actors to evade accountability through cross-border regulatory arbitrage.

3.2.3 Establish adaptive regulatory mechanisms

To ensure that legal frameworks remain fit for purpose in the face of rapidly evolving AI capabilities, adaptability must be embedded into the very structure of AI governance. Static rules are unlikely to withstand the pace and complexity of innovation in AI-generated content and algorithmic manipulation. Instead, a dynamic, data-driven, and iterative regulatory architecture is needed. One effective approach is the introduction of regulatory sandboxes tailored to AI applications in the information environment. These sandboxes provide controlled environments where platforms, regulators, and civil society actors can test and refine new moderation or detection technologies before they are mandated at scale. For instance, a platform might deploy a prototype AI labeler for likely disinformation under regulatory oversight, allowing for real-time learning and adjustment. Such experimentation can inform key design questions, such as acceptable error margins, redress mechanisms for false positives, and proportionality thresholds for intervention. These insights are crucial before formalizing policies into law.

In parallel, the inclusion of sunset clauses and scheduled policy reviews should become a standard feature of AI-related legislation. Laws passed today may be obsolete within two years; thus, built-in review cycles (e.g., every 18–24 months) allow regulatory frameworks to evolve alongside technological progress. In addition, legal mechanisms should empower relevant authorities to trigger accelerated updates in response to breakthroughs or emergent threats – such as audio deepfakes used in impersonation scams or novel bot networks designed to hijack trending topics. We further propose the formation of dedicated AI threat foresight and monitoring units at the national and regional levels. These bodies would be tasked with horizon scanning for emergent disinformation vectors, maintaining ongoing risk assessments, and coordinating with global partners for early warning and preemptive guidance. Such units could operate under the auspices of broader regulatory agencies or be integrated into the proposed Global AI Security Alliance. Importantly, adaptability does not mean deregulation. It requires a clear procedural framework to evaluate when and how policies should be revised – grounded in empirical evidence, ethical principles, and inclusive stakeholder dialogue.

This adaptive approach acknowledges that disinformation is a moving target and must be governed as a continuously evolving socio-technical risk, not a one-time legislative challenge. By embracing regulatory agility, policymakers can strike a balance between fostering innovation and safeguarding democratic discourse in an age of algorithmic manipulation.

3.2.4 Facilitate agile, pre-competitive collaboration to develop adaptive defenses

Technology is not only part of the problem – it is also an essential part of the solution. Rather than relying solely on isolated innovation or reactive investment, we recommend fostering agile, pre-competitive collaboration among public, private, and academic stakeholders to accelerate the development of adaptive defenses against AI-enabled disinformation. Governments, platforms, and international bodies should jointly fund interdisciplinary R&D initiatives at the intersection of AI, cybersecurity, behavioral science, and information integrity. This includes building shared threat modeling environments, where researchers can test adversarial AI tactics and co-develop countermeasures. For instance, new generative detection models could analyze imperceptible anomalies in audio, image, or video artifacts left behind by synthetic content – an approach already showing promise in deepfake identification. AI-powered verification tools are also essential. Leveraging real-time cross-referencing with trusted sources and metadata can help determine the authenticity of viral content. Equally critical is cognitive resilience – developing tools and educational interventions to help users recognize and resist manipulation. Adaptive AI systems can deliver timely "prebunking" prompts or context-aware alerts to users exposed to potentially misleading content, based on risk indicators such as origin, format, or linguistic style.

Complementary technologies such as blockchain-based provenance tracking and cryptographic watermarking offer innovative tools for bridging these gaps. Blockchain can enhance content traceability by recording immutable, timestamped hashes and metadata at the point of media creation – providing tamper-evident provenance that persists even as content is shared and modified. Projects like the Content Authenticity Initiative and Project Origin demonstrate how these tools could be embedded at scale. Beyond content traceability, blockchain can support AI accountability mechanisms. Developers and providers of generative AI tools could collaborate through decentralized autonomous organizations (DAOs) or self-regulatory bodies, using blockchain-backed smart contracts to formalize shared norms and compliance mechanisms to implement smart contracts that codify ethical deployment commitments. These may include obligations to watermark AI outputs, audit downstream usage, or enable redress mechanisms when misuse occurs. Such on-chain mechanisms create transparent, verifiable records without compromising user privacy. Importantly, blockchain's decentralized, immutable nature aligns with the need for distributed oversight of powerful AI systems. While challenges remain – such as transaction scalability, energy use, and integration with legal frameworks – these technologies offer a blueprint for tamper-resistant infrastructure that complements existing governance proposals. Policymakers should invest in pilot programs and regulatory sandboxes to explore how blockchain-based mechanisms can be embedded in future AI oversight architectures.

3.2.5 Promote digital literacy and cognitive resilience

Technological and policy measures must be complemented by efforts to strengthen the human element of resilience. A well-informed and vigilant public is one of the best defenses against disinformation. We recommend national and international initiatives to educate citizens about AI-generated content and online manipulation techniques. This should start in schools by integrating media literacy and critical thinking into curricula, including specific modules on deepfakes, synthetic media, and the tricks used in viral misinformation. Already, several countries and NGOs have piloted programs to teach students how to recognize false or manipulated content; these should be expanded and shared globally. Public awareness campaigns are also needed for the adult population. Governments, in collaboration with civil society and media organizations, can run information drives on social media and traditional media, illustrating common examples of AI-fabricated news and how to spot them. Platforms can assist by providing easy-to-use tools for users to fact-check content or trace its source with one click (for example, plug-ins that reveal an image's origin or whether a video has been edited). Libraries, community centers, and workplaces could host workshops on navigating misinformation online. The overarching goal is to build cognitive resilience – the mental ability to critically assess information and resist manipulation. Researchers describe cognitive resilience as akin to a *"cognitive firewall"* that prevents false information from taking root (Kont et al., 2024). This can be cultivated through "prebunking" and inoculation strategies: for example, exposing people to weakened doses of common misinformation tropes and debunking techniques so that they are less susceptible when they encounter falsehoods in the wild (van der Linden et al., 2021). Platforms might deploy brief pop-up warnings or tutorials for users who are about to share content that has characteristics of a deepfake or bot-originated post, thereby nudging users to pause and verify. Over time, a more discerning public will reduce the effectiveness of disinformation, as false narratives fail to gain traction and credibility. While digital/media literacy alone cannot stop a determined influence campaign, it *raises the costs* for disinformers and can mitigate the damage. Importantly, these efforts also contribute to healthier civic discourse by encouraging people to seek reliable sources and engage critically rather than impulsively with provocative content.

In implementing these recommendations, a phased approach can be useful. Some actions (like bias audits and tiered regulations for small businesses, or launching literacy campaigns) can be taken immediately as "short-term" measures, as they have moderate impact but lay the groundwork. Transitional steps (over the next 1–2 years) include establishing cooperation frameworks and adaptive regulatory processes, which have higher impact and need careful planning. Finally, long-term strategies (3–5 years and beyond) such as large-scale research initiatives and global governance agreements will yield the highest impact in fortifying the information ecosystem. By combining quick wins with sustained strategic efforts, governments and stakeholders can progressively reinforce society's defenses.

4 Discussions of cases of with role of AI used to produce or to detect disinformation that we studied

4.1 Classification of types of AI and their roles in disinformation

To illustrate the interplay between AI technologies, disinformation tactics, and societal impacts, we examine recent cases and developments across several categories of AI application. Below, we classify the main ways AI is used to produce disinformation and

provide real-world examples of each, along with the harms they have caused. We also discuss how AI is being employed in counter-disinformation efforts. This classification underscores the dual nature of AI – as a tool for malign manipulation and as an instrument for defense.

AI techniques leveraged in disinformation campaigns can be grouped into a few broad categories:

Generative AI for content creation – this includes *deepfakes* (AI-generated synthetic video or audio that imitates real people) and *AI-generated text* (using large language models). These tools can fabricate convincing false content at scale.

Synthetic identities and personas – AI can create realistic profile pictures, names, and personal backgrounds, enabling the mass production of fictitious online personas that appear authentic.

Automation and bot networks – AI-driven bots can mimic human behavior on social media, posting and engaging with content automatically. Machine learning helps coordinate these bots to act in swarms or networks.

Coordination of disinformation at scale – real-time social media analysis and strategy adaptation, cross-platform botnet deployment, natural language generation in multiple languages and styles.

AI in counter-disinformation – on the defensive side, AI is used to detect fake content and accounts, as well as to automate fact-checking and content moderation.

We now delve into each of the offensive categories (first four) with case studies, and then discuss the defensive use of AI.

4.2 Generative AI for content creation (deepfakes)

AI-generated synthetic media, or deepfakes, have become one of the most visible tools of disinformation. Deepfake detection company Onfido reported a 3,000% increase in deepfake attempts in 2023 (Onfido, 2023). The global market value for AI-generated deepfakes is projected to reach $79.1 million by the end of 2024, with a compound annual growth rate (CAGR) of 37.6% (MarketsandMarkets, 2024). There were 95,820 deepfake videos in 2023 – a 550% increase since 2019, with the number doubling approximately every six months (MarketsandMarkets, 2024). In 2023, deepfake fraud attempts accounted for 6.5% of total fraud incidents, marking a 2,137% increase over the past three years (Onfido, 2023).

Deepfakes utilize advanced neural networks to create fake video or audio that is often difficult to distinguish from real recordings, thereby manipulating viewers' perceptions of reality. A number of high-profile incidents in recent years demonstrate their disruptive potential:

4.2.1 Political deception

During the Russian invasion of Ukraine, a fabricated video emerged online in March 2022 that depicted Ukrainian President Volodymyr Zelenskyy apparently urging his troops to lay down their arms and surrender. This video, a deepfake, was cleverly edited into a social media broadcast format and spread rapidly on Facebook and other platforms before being debunked. Had it been widely believed, it could have eroded military morale and undermined public trust in official communications during a critical national crisis. Ukrainian news outlets and government spokespeople rushed to clarify that the video was fake, illustrating both the risk and the swift response such disinformation provokes (U.S. Department of State, 2024).

4.2.2 Election interference

In the 2024 Indian regional elections, deepfake videos were deployed to subvert language barriers and spread targeted propaganda. Reports indicate that partisan operatives created videos of political leaders mouthing words in languages they never spoke, tailoring messages to different ethnic groups. Over 15 million people were reached via about 5,800 WhatsApp groups that circulated these deepfakes (Election Commission of India, 2024). By exploiting India's linguistic diversity, the campaign manipulated voter perceptions of rival candidates. This not only misled voters on the politicians' actual statements but also inflamed tensions via disinformation that appeared to come from trusted community figures. Election officials noted that such tactics, if unchecked, could compromise electoral integrity and make it harder for voters to discern truth amid a flood of AI-fabricated content.

4.2.3 Stoking public distrust

In Slovakia, on the eve of the 2024 parliamentary elections, an AI-generated audio recording surfaced in which voices resembling two prominent politicians discussed plans to rig the election (Slovak Ministry of Interior Ministry of the Interior of the Slovak Republic, 2024). Even though the audio was quickly suspected to be fake, it circulated widely on messaging apps. The content of the deepfake phone call fed into existing anxieties about corruption, leading some citizens to question the legitimacy of the election process. This incident eroded trust in the democratic process; officials later confirmed no such conversation had occurred, but the damage was done in terms of sowing doubt. Voter turnout in certain areas was thought to be depressed by fears that the system was rigged, showing how a simple audio deepfake can have tangible effects on civic behavior.

4.2.4 Personal attacks and intimidation

Deepfakes have also been weaponized to harass individuals and attempt to silence voices. In Northern Ireland, Member of Parliament Cara Hunter was targeted in 2023 by a malicious deepfake pornography campaign (Hunter C., 2024). Someone used AI to graft her likeness onto explicit sexual content and disseminated the fake video online. The reputational damage and emotional trauma from this incident were significant. Beyond the personal toll on Ms. Hunter, observers feared such tactics could discourage women and young people from participating in politics, if they see that outspoken figures risk being humiliated by fabricated scandals. This case drew attention to the need for stronger legal recourse against deepfake harassment, and indeed, the UK Parliament cited it in debates on criminalizing certain uses of deepfakes.

These examples highlight the various social and political harms deepfakes can inflict: from confusion on the battlefield, to manipulated democratic decisions, to the chilling effect on public life. The technological advancement in deepfake quality is rapid. Recent statistics show a three-fold increase in the number of deepfake videos and an eight-fold increase in deepfake audio clips circulating online

from 2022 to 2023. By 2023, an estimated 500,000 deepfake videos were shared on social media. As deepfakes become more common and harder to detect, the necessity of robust countermeasures grows. Efforts are underway on multiple fronts: researchers are developing AI-driven deepfake detection tools, legislators in several countries are drafting laws to penalize malicious deepfake creation (especially in contexts like election interference or defamation), and media literacy campaigns are educating the public to be skeptical of sensational video/audio clips. These responses are critical. *If* deepfakes can be exposed and contextualized quickly, their impact can be mitigated. A combination of technical detection (using algorithms to verify authentic media or spot anomalies) and comprehensive media literacy is seen as the best defense. In summary, deepfakes represent a significant new dimension of disinformation – one that directly targets our sense of reality – and combating them will remain a top priority for policymakers and technologists alike.

4.3 Generative AI for content creation (AI-generated text)

Large Language Models (LLMs) such as GPT-3 and GPT-4 have enabled the mass production of synthetic text that is often highly coherent and hard to distinguish from human writing. This capability has been co-opted by disinformation actors to flood social media and news sites with fabricated articles, posts, and comments, exploiting the scale and speed that AI text generation affords. There are several dimensions to this phenomenon:

On the supply side, the accessibility of powerful text generators has grown. The global market for AI text generation tools was estimated at $423.8 million in 2022 and is projected to reach $2.2 billion by 2032, expanding at a rapid rate as businesses and individuals adopt these models for various uses (Dergipark, 2024). This growth means the tools are widely available and becoming cheaper, lowering the barrier for malicious use. North America currently leads in usage share, but Asia-Pacific is expected to see the fastest growth, indicating a geographically widening usage (Grand View Research, 2024). As more actors gain access to these AI systems, the potential volume of AI-written disinformation increases correspondingly.

We have witnessed misinformation campaigns driven by AI text causing real-world impacts. During the COVID-19 pandemic, for example, numerous false narratives about vaccines and health measures were propagated through what appeared to be legitimate news articles and blog posts. Investigations later revealed that some of these pieces were authored by AI systems and then posted on websites masquerading as news outlets, or shared via social media bots. These AI-generated articles made baseless claims (such as exaggerating vaccine side effects or promoting fake cures) but were written in a convincing journalistic style. According to the World Economic Forum, such health misinformation contributed to increased vaccine hesitancy in multiple countries. The public, already fearful due to the pandemic, encountered what looked like factual reports, not realizing they were computational concoctions. This demonstrates how AI-generated text can amplify the reach of harmful falsehoods by sheer quantity and the illusion of legitimacy, thereby undermining public trust in health authorities and complicating crisis responses.

In the political realm, AI text generators have been harnessed to produce persuasive messaging at a volume and personalization level not achievable before. Political consultants and propagandists can use tools like GPT-based text generation platforms to draft tailored emails, manifestos, or social media posts targeting specific voter demographics. One reported case involved the use of a tool called *Quiller.ai* during recent elections to help a particular campaign draft thousands of unique fundraising emails and social media posts. The AI was given basic points and the profiles of target recipients, and it produced content that resonated with those individuals' known interests and fears. While fundraising itself is legitimate, blending this strategy with propaganda crosses into disinformation when the messages contain misleading claims or emotionally manipulative rhetoric disconnected from facts. The use of AI blurred the lines between genuine grassroots communication and mass-produced propaganda, making it harder for recipients to tell if a heartfelt plea on social media was written by a real supporter or generated by an algorithm. This raises ethical questions about authenticity in political discourse and shows how AI can supercharge microtargeting efforts with minimal human effort.

Another telling example comes from the 2016 Brexit referendum in the UK, which, although predating the latest AI advances, foreshadowed tactics that AI is now amplifying. Campaigners on both sides employed microtargeted advertising on Facebook to deliver custom messages to voters based on their data profiles. According to subsequent analyses, many of these messages were misleading or fear-mongering. Fast forward to today: similar microtargeting can be conducted by AI agents autonomously generating content. Reports indicate that in follow-up campaigns and discussions around Brexit, *automated persona accounts* (some using AI-generated profile photos and AI-written posts) engaged UK voters by pushing emotionally charged narratives – such as exaggerated fears about immigration or economic doom – and these were tailored to individuals' online behavior patterns. The AI essentially acted as a propagandist that learned what each segment of the population cared about and then produced slogans and "news" addressing those exact fears. The concern is that such personalized disinformation is far more convincing than one-size-fits-all falsehoods; it can quietly reinforce people's biases and is difficult to challenge because each person may see a slightly different misleading message, hidden from public scrutiny.

The economic impacts of AI-generated text-based disinformation are non-trivial. One analysis by the World Economic Forum in 2020 estimated global economic losses of around $78 billion in that year due to misinformation and fake news spreading online. These losses come from various channels: scams and frauds (often enabled by fake emails or news that trick people into financial decisions), companies losing value due to false rumors, resources spent on debunking hoaxes, and broader erosion of trust in markets and institutions. If AI allows misinformation to scale up, these economic costs could grow. We have already seen stock prices dip or surge based on viral social media claims – some notable cases involved automated Twitter bots spreading false reports about companies, causing brief chaos in financial markets before corrections. As LLMs become integrated into bots, the false reports could become more elaborate and harder to immediately dismiss, potentially leading to more severe market manipulation incidents.

Public perception data underscores the seriousness of the challenge. Surveys show that large portions of the population are aware of and worried about AI's role in creating misinformation. In the UK, 75% of adults in 2023 believed that digitally altered videos and images (e.g., deepfakes) contributed to the spread of online misinformation, and 67%

felt that AI-generated content of all types was making it harder to tell truth from falsehood on the internet (Ofcom, 2023). Globally, more than 60% of news consumers believe that news organizations at least occasionally report stories they know to be false, a cynicism fueled in part by awareness of mis/disinformation dynamics. Intriguingly, 38.2% of U.S. news consumers admit to having unknowingly shared a fake news item on social media, only realizing later that it was false. These figures highlight a growing distrust in media and the self-reinforcing nature of misinformation – people are both victims and unwitting vectors of disinformation in the online ecosystem. Journalists themselves are highly concerned: 94% of journalists surveyed see made-up news as a significant problem in their field. This situation creates a vicious cycle where disinformation, boosted by AI, begets more distrust, which in turn primes the public to be more susceptible to the next wave of disinformation or to dismiss truthful reporting as "fake."

In conclusion, AI text generation tools have become double-edged swords. They offer efficiency and creativity but in the wrong hands can greatly magnify the reach and believability of false information. Addressing this will require not only better detection algorithms (to flag AI-written trolls or bogus news) but also platform policies to throttle or label automated accounts, and a culture of critical thinking among readers. Some social networks are exploring authenticity verification for accounts and limiting bot activity, while researchers are developing methods to watermark AI-generated text to aid detection. However, adversarial use of AI will likely circumvent simpler safeguards, meaning the guardians of information integrity will need to continuously adapt, perhaps even employing *counter-LLMs* that identify linguistic patterns of AI vs. human text. The battle between AI-generated disinformation and AI-enabled detection is already underway as a key front in maintaining a healthy information space.

4.4 Synthetic identities and personas

Another troubling aspect of AI in disinformation is the creation of synthetic identities – fake personas generated with the help of AI to appear as real people online. These typically involve AI-generated profile pictures (often using generative adversarial networks to create realistic human faces that do not correspond to any actual person), along with fabricated names and life details. Networks of such personas can then be used to amplify false narratives, infiltrate groups, or lend false credibility to disinformation by posing as "concerned citizens" or experts. The use of synthetic identities has grown in both criminal fraud and political propaganda. Here we detail some notable instances:

One early domain of synthetic identity abuse was financial fraud. A nationwide fraud ring in the United States around 2020–2021 demonstrated how blending real and fake information could dupe financial institutions. The perpetrators created synthetic identities by pairing fabricated personal details (names, addresses) with real Social Security numbers stolen from individuals (often children who would not be using their SSNs yet). With these identities, they opened bank accounts, obtained credit, and even set up shell companies, all under fictitious personas. Over time, they built credit histories for these "people" and then defaulted on loans and credit lines, stealing nearly $2 million from banks before being caught (Department of Homeland Security, 2024). While this case was primarily a financial crime, it highlighted the ease of creating credible fake personas in the digital record-keeping systems. Once the concept proved effective for money theft, it was only a matter of time before similar techniques were applied to disinformation efforts – where the currency is influence rather than cash.

Indeed, by 2023 the use of AI-generated profile pictures and personas had become a common feature of online propaganda campaigns. In a striking example, Meta (Facebook's parent company) reported taking down a network of nearly 1,000 fake accounts across Facebook and Instagram that had AI-generated profile images and elaborate backstories. These accounts posed as a diverse array of individuals: some pretended to be protest activists, others were posing as journalists or young women (Meta, 2023). In reality, all were controlled by an organization pushing pro-authoritarian narratives in various countries. This operation – publicly revealed in Meta's (2023) Coordinated Inauthentic Behavior report – demonstrated how AI can dramatically scale up "troll farms." Instead of stealing profile photos or reusing images (which reverse-image search can detect), the operators used generative adversarial networks to create unique faces, making it much harder for automated systems to flag them as fake. By blending these synthetic personas into genuine online communities, the campaign was able to more credibly insert false stories and extremist talking points into public discourse, as the messages appeared to be coming from ordinary people rather than overt bots or known sockpuppet accounts.

During the COVID-19 pandemic, public health misinformation provided another arena for synthetic identities. Anti-vaccine groups and conspiratorial movements employed AI-generated profiles on Twitter, Facebook, and fringe platforms to bolster their ranks. For instance, an investigation by the Centers for Disease Control and Prevention (CDC) (2021) found that hundreds of social media profiles spreading false claims about vaccines (such as extreme exaggerations of side effects, microchip myths, etc.) were not real people, even though their profile photos and names looked authentic (Centers for Disease Control and Prevention, 2020–2021). These synthetic influencers amassed followers and engaged with real users, helping to spread false information widely and making it seem as if there was a larger grassroots community doubting vaccines than actually existed.

While these cases highlight the dangers of synthetic identity misuse, it is essential to distinguish between the underlying technology and its application. Generative AI tools that create convincing profiles can also be used by defenders – for instance, to train detection systems, simulate threat scenarios, or build honeypot accounts to trace malicious networks.

Synthetic identities have also been weaponized in the political context of elections and referendums. A retrospective look at the 2016 Brexit campaign indicated the presence of suspicious accounts later suspected to be fictitious. But in more recent electoral events, concrete evidence has emerged. In the lead-up to various elections, including the 2020 U.S. elections and others, researchers documented fake persona accounts used to deliver tailored political messages. These messages exploited voters' fears and preferences to influence their decisions. In the Brexit context, it has been reported that some AI-driven profiles posed as British citizens on social media and delivered customized propaganda – for example, sending anti-immigrant scare stories to voters identified (through data analysis) as concerned about immigration, or sending exaggerated economic doom stories to those worried about EU regulations. By doing so in a targeted and anonymous manner, these synthetic actors could sway

opinions while evading immediate detection. The UK Electoral Commission's inquiry into digital campaigning noted such activity as a harbinger of future threats to fair democratic debate (UK Electoral Commission, 2016).

Beyond text-based influence, synthetic identities combined with deepfakes have facilitated sophisticated impersonation scams. In one particularly brazen case, cybercriminals created a deepfake video of the Chief Communications Officer of the cryptocurrency exchange Binance (Patrick Hillmann) and used it to trick representatives of various crypto projects. The scammers, impersonating a top executive via a lifelike AI-generated video avatar on Zoom calls, convinced people that Binance would list their cryptocurrency tokens in exchange for a fee. This incident not only defrauded victims of money but also demonstrated the fusion of synthetic identity and deepfake technology for financial gain. It shows that synthetic identities are not limited to static profiles; they can extend to real-time avatars that engage interactively. The fact that even savvy crypto entrepreneurs fell for the ruse underscores the challenge – visual evidence that traditionally would confirm identity (a face-to-face video meeting) can no longer be taken at face value in the age of AI (Hillmann, 2023).

Synthetic identities have been used at high levels of political deception as well. In 2022, the mayors of Berlin, Madrid, and Vienna were each, on separate occasions, tricked into holding video calls with someone they believed to be Vitali Klitschko, the Mayor of Kyiv. In reality, they were speaking to an imposter using a deepfake of Klitschko's face and voice. The imposter (whose true identity remains unclear, though suspicion fell on Russian actors) raised controversial issues in these calls – such as asking about sending Ukrainian refugees back home to fight – apparently in an attempt to elicit embarrassing or divisive responses from the European mayors. While these particular mayors quickly grew suspicious and the calls were cut short, the incident revealed how far synthetic identity deception can go: reaching directly into high-level diplomatic conversations. The European Parliament later discussed this incident as a warning of the potential for diplomatic interference using AI. It highlighted the urgent need for verification protocols in video communications; for example, employing code words or secondary channels to confirm identities during sensitive discussions (European Parliament, 2022).

There are even instances where state-run media themselves deploy synthetic personas as part of their propaganda. It has been alleged, for example, that Chinese state media created an entirely fictional Italian man (complete with an AI-generated face and a social media presence) who appeared in a video thanking China for its aid during the COVID-19 pandemic. This video was circulated by Chinese outlets to demonstrate international appreciation for China's efforts, but investigators found that the "Italian" individuals shown were likely deepfakes or actors, not real citizens spontaneously expressing gratitude. While not a direct attack, this is state-sponsored narrative reinforcement using synthetic means, demonstrating that not only shadowy groups but also governments may use AI-created people to serve their messaging goals (China Global Television Network, 2020).

Just as synthetic personas have been used to deceive, the same AI techniques can be harnessed for defense. Detection algorithms trained on large datasets of fake and real profiles are now capable of spotting GAN-generated faces with increasingly high precision. Behavioral pattern recognition – such as unusually rapid group joining, repetitive posting styles, or cross-platform profile reuse – can flag inauthentic activity at scale. Adversarial AI models can be deployed to test the robustness of social media filters, helping platforms identify vulnerabilities before real attackers do. These positive use cases underscore that it is not the AI itself that is harmful, but rather how it is used – an insight that must anchor any regulatory or policy response.

In summary, synthetic identities represent a powerful tactic in the disinformation playbook. They enable one operative to appear as many voices, create the illusion of consensus or grassroots support, infiltrate groups under false pretenses, and even impersonate trusted individuals. The harm from these cases ranges from financial losses and reputational damage to skewing democratic discourse and international relations. Combating synthetic identities is challenging: it requires improved authentication methods (for example, social media companies deploying verification badges or using AI to detect when profile pictures are AI-generated), as well as user vigilance. Some progress has been made – for instance, researchers can sometimes identify GAN-generated profile images through subtle telltale signs (like anomalies in the rendering of ears or jewelry), and social platforms have begun using algorithms to flag faces that appear too similar to known AI face datasets. However, as AI improves, these fakes will become more indistinguishable.

Thus, detection efforts must continually evolve, and platforms should consider limiting certain behaviors (like new accounts rapidly joining hundreds of groups) that synthetic personas often engage in. Policy responses might include requiring disclosure of the use of AI in generating profile content (an element found in the EU's draft AI Act, mandating labeling of AI-generated media). Ultimately, diminishing the influence of synthetic identities will involve both technical defenses and a degree of digital skepticism among users – verifying identities through multiple sources when claims from "people" online seem extraordinary.

4.5 Automation and bot networks

Automated social media accounts – or bots – have long been a feature of online platforms, but AI has made them far more sophisticated and effective agents of disinformation. Modern AI-driven bots can generate original content, engage in conversations, and operate in coordinated networks that amplify false narratives. They effectively act as force multipliers for propagandists: each human operator can deploy a legion of bot accounts to do the work of spreading and magnifying messages. Two cases illustrate the capabilities and tactics of AI-driven bot networks, followed by a discussion of their broader implications.

4.5.1 Case 1: Russian AI-enhanced disinformation network (2024)

In July 2024, U. S. authorities exposed a large-scale Russian disinformation operation that had leveraged a custom AI software tool called "Meliorator." This tool was used by Russia's intelligence services to create and manage a fleet of fictitious social media personas on Twitter (recently rebranded as X) and potentially other platforms. Nearly 1,000 fake accounts were identified as part of this network. What set them apart was the level of detail: these AI-generated accounts did not have the typical markers of bots (like random usernames or lack of personal info). Instead, each had a realistic profile picture (courtesy of AI image generation), a plausible name, and even a backstory including political views and local affiliations. The bots operated by posting pro-Kremlin narratives about the war in

Ukraine and other geopolitical issues, mixing these posts into trending conversations to increase visibility. They would follow real users, like and retweet content, and even respond in threads with arguments favoring Russia's position, thereby mimicking human behavior to a high degree (Department of Homeland Security, 2024).

The scale and coordination were key. Meliorator allowed a small team to orchestrate this thousand-strong army of bots, timing their posts and interactions to maximize impact. For example, when a piece of news unfavorable to Russia emerged, dozens of these fake personas would swiftly respond or post counter-messaging to seed doubt or alternative interpretations. The campaign aimed to undermine support for Ukraine among Western audiences and to bolster narratives favorable to the Russian government. U.S. investigators noted that the operation blurred the lines between authentic grassroots commentary and state-sponsored propaganda, thereby polluting the information environment around the Ukraine conflict. It took significant analytical effort (including AI-based detection tools on the defenders' side) to identify these accounts as fake. Telltale patterns – such as metadata similarities and coordinated posting times – eventually revealed the network. This case exemplifies how a hostile actor can use AI to wage information warfare covertly, and it underscores the importance of continually improving detection techniques and international cooperation to counter such threats.

4.5.2 Case 2: "Reopen America" COVID protests (2020)

During April 2020, as the COVID-19 pandemic led to lockdowns and public health restrictions, a sudden wave of protests erupted in various U. S. states under the banner "Reopen America." Subsequent analysis of social media data suggested that an orchestrated bot campaign helped spark and amplify these protests. Researchers at the University of Southern California examined Twitter traffic during that period and found a network of automated accounts that were synchronizing hashtags and slogans across platforms (University of Southern California, 2020). These bots, likely guided by AI, would all tweet messages like "#Reopen [State]" and calls to end lockdown, often at the same timestamps and with identical wording, indicating a high level of coordination. By doing so, they managed to push these hashtags into Twitter's trending topics, which in turn gave the movement more visibility and a perception of momentum.

The illusion of widespread support created by the bots had real consequences. Seeing "Reopen" trends and large numbers of social media posts advocating against lockdowns, more individuals joined in, both online and in physical protests. It was later noted that many legitimate users were interacting with or following these bot accounts, thinking they were fellow concerned citizens. The coordinated bot activity amplified contentious narratives, such as framing public health measures as tyranny or emphasizing fringe conspiracy theories about the virus, thereby polarizing public opinion further. Investigations hinted that some of these operations could have been backed by partisan groups or foreign entities aiming to sow chaos. The net effect was that what might have been isolated local dissent was inflated into a national movement. The campaign contributed to actual rallies in state capitals, with crowds – fueled by misinformation about COVID – demanding the lifting of restrictions, potentially undermining the pandemic response and endangering public health. This case shows that even in democratic societies, bots can significantly distort the public square, driving events in the physical world by manipulating narratives in the virtual one.

4.5.3 Case 3: Stockport-related riots (UK, 2023)

Localized riots were triggered by a fabricated story alleging violence involving asylum seekers. The story, initially shared via encrypted messaging apps, was amplified by social media bots and fueled real-world unrest. Investigations later found that parts of the story originated from AI-generated sources, and synthetic identities were involved in the online amplification. The case shows how AI-enhanced, bot-driven disinformation – even at the local level – can rapidly incite real-world violence when detection and response mechanisms are absent or delayed (BBC News, 2023; Full Fact, 2023; Norton and Marchal, 2023).

4.5.4 Case 4: AI-powered disinformation in China's "Spamouflage" campaign (2022–2023)

A striking example of AI-driven automation is the Chinese state-aligned "Spamouflage" operation. This campaign leveraged bots and generative AI in tandem to disseminate propaganda and interfere in foreign elections (Stanford Internet Observatory, 2024; Meta, 2023; U.S. Department of State, 2024). It featured: AI-generated deepfake video anchors on a fictitious outlet, *Wolf News*, delivering pro-China narratives in English using synthesized avatars created with tools like Synthesia (Stanford Internet Observatory, 2024); Synthetic audio deepfakes, including a falsified voice clip of Taiwanese candidate Terry Gou endorsing a rival – an attempt to meddle in Taiwan's 2024 election (U.S. Department of State, 2024); AI-crafted memes and manipulated images promoting anti-US and anti-Japan messages (Meta, 2023); Coordination through bot networks to amplify all of the above (Stanford Internet Observatory, 2024). This campaign exemplifies the evolving sophistication of AI-driven disinformation. Instead of merely automating reposts, AI now creates original persuasive content – raising the ceiling of what bots can achieve in propaganda ecosystems. Detection efforts highlighted both the promise and the present limitations of deepfake forensics: while current-generation avatars showed tell-tale signs like robotic intonation and visual glitches, future iterations may become indistinguishable from authentic media. The Spamouflage case illustrates a turning point: disinformation actors now pair generative tools with automation to conduct scalable, persuasive, and potentially election-disrupting campaigns (Stanford Internet Observatory, 2024; Meta, 2023).

The broader implications of such AI-driven bot networks are far-reaching:

4.5.5 Erosion of trust

As bots become more prevalent and harder to distinguish from real users, people grow more skeptical about online interactions and content. The phenomenon known as the "liar's dividend" becomes a risk, where even truthful content is questioned because the public knows fakes are out there. When any opposing voice or inconvenient fact can be dismissed as "just a bot" or a deepfake, genuine democratic discourse suffers. For instance, authentic grassroots campaigns might struggle to gain credibility if observers suspect that social media support could be manufactured. UNESCO and Ipsos reported in 2024 that a majority of internet users worldwide feel they cannot be sure if what they see on social media is real or manipulated, indicating a

general environment of distrust that is exacerbated by bot-driven disinformation (Ipsos and UNESCO, 2024).

4.5.6 Manipulation of public opinion

AI-driven bots can manufacture consensus or at least the appearance of it. By making a particular viewpoint or hashtag trend, they can persuade neutral observers that "everyone is talking about this" or that a certain extreme position is more widely held than it truly is. This can attract media coverage (the press often reports on trending topics), further amplifying the desired narrative. For instance, during the "Reopen America" protests, the coordinated network made it seem as if there was a broad grassroots uprising against lockdowns, possibly influencing policymakers to consider lifting restrictions earlier than some health experts advised. Bots also engage in dogpiling – swarming individuals who voice opposing views with waves of criticism or harassment, sometimes driving them off the platform and silencing their perspective. In authoritarian contexts, regime-linked bots have been used to drown out dissenting hashtags by overwhelming them with pro-regime messages, effectively smothering opposition voices online.

4.5.7 Information overload and "noise"

The ability of bots to produce content at a much higher volume than humans leads to an asymmetric flood of information. Studies have quantified this: one study found that bots can be 66 times more active than ordinary users, and in certain contentious discussions, bots made up nearly one-third of the content despite being a tiny fraction (under 1%) of participants (Rossetti and Zaman, 2023). This deluge can crowd out factual information. During breaking news events, for example, bot accounts might rapidly push misinformation or distracting content, complicating the job of journalists and fact-checkers to identify what's true. For the average user, trying to sift through an avalanche of posts – some human, many automated – creates fatigue and confusion, which disinformation campaigns can exploit (people might start to believe a falsehood simply because they have seen it so many times in their feed, a familiarity effect).

4.5.8 Amplification of low-credibility sources

Bots often serve to aggressively promote links from fringe websites or known propaganda outlets. Research shows that a significant portion of shares for content from "low-credibility" sources (those that regularly publish false or misleading info) can be attributed to bots. One study in 2024 indicated that about 33% of the top sharers of articles from dubious sites were likely automated accounts (George Washington University, 2024). This means bots can help such content leap into the mainstream conversation, whereas it might have languished in obscurity otherwise. By artificially boosting the hit counts and share counts of these articles, bots can even game platform algorithms that elevate content based on engagement metrics. This amplification is especially pernicious during elections or referendums, as false narratives from fringe outlets can get enough visibility to potentially sway undecided voters or reinforce polarized views.

4.5.9 Undermining democratic processes

The presence of swarms of bots in political communication has raised alarms about electoral integrity. In the 2016 U.S. presidential election, analysts estimate that roughly 19% of tweets about the election came from bot accounts (Akamai, 2024). These accounts were found to be disproportionately pushing either extreme viewpoints or supportive messages for particular candidates, thus skewing the online discourse. In close races, the influence of bots – by driving agendas, altering perceptions of candidate popularity, or spreading disinformation about voting procedures – could even be outcome-determinative. More subtly, the knowledge that bots are interfering can reduce public confidence in election results; people might suspect the "voice of the people" was tainted by fake participants. All of this undermines the legitimacy of democratic outcomes and can contribute to unrest or refusal to accept results, as seen in various disputes where online disinfo played a role in fueling skepticism.

In response to the bot challenge, social media companies have tried to crack down by improving bot detection and removal. For example, Twitter (prior to its rebranding) routinely eliminated millions of suspected bot accounts every quarter and introduced rate limits to prevent excessive posting. However, the adversaries adapt as well – using techniques like time-shifting (to post during normal human hours) and content variation (to avoid obvious copy-paste patterns) to slip past filters. The cat-and-mouse game between platform moderators and bot creators is ongoing. On the policy side, some jurisdictions are considering or have passed laws requiring disclosure if content is produced by a bot (California enacted a B.O.T. act for certain commercial and political bots). But enforcement is tricky across borders.

One promising angle is the development of network analysis tools that look not just at individual accounts but at their behavior in aggregate, identifying telltale signs of coordinated campaigns. The 2024 Russian Meliorator network was discovered through such analysis, finding the connective tissue between accounts. Academic and industry researchers are increasingly collaborating to map out these networks (e.g., the Computational Propaganda Project at George Washington University cataloguing global bot operations).

Ultimately, tackling malicious bots may also require verification and authenticity standards on social media – such as optional verified identity for users (with trade-offs in terms of anonymity and privacy debates). If users can filter to see only verified humans, that could reduce bot impact, though it raises equity issues (not everyone has ID or wants to disclose it).

In conclusion, AI-driven bots are a cornerstone of contemporary disinformation efforts, capable of distorting online discourse at an unprecedented scale. They exploit the openness of social platforms and the psychological biases of users. Strengthening our defenses against them is critical to maintaining a functional digital public sphere.

4.6 Coordination at scale

The synergy of AI techniques across content creation, automation, and targeting enables large-scale coordinated disinformation campaigns that cut across platforms and national boundaries. Rather than isolated tactics, we now see integrated operations where AI monitors the information ecosystem, generates tailored content, and disseminates it through a web of channels for maximum impact. Key

aspects and statistics illustrate how this coordination works and its effects:

Real-time social media analysis and strategy adaptation – disinformation campaigns are employing advanced AI-driven analytics to continuously scan social media trends, viral content, and user sentiment. By doing so, propagandists can identify emerging narratives or breaking news events to exploit. For example, if a natural disaster strikes, an AI system might quickly gauge public fears and then suggest disinformation angles (like conspiracies or blame narratives) to push. These systems can also detect when their false narratives are gaining or losing traction and adjust accordingly. This feedback loop allows for agile, adaptive tactics in almost real time. For instance, during an election, if a particular misleading talking point is not resonating, the AI may pivot to another line of attack that data shows is more compelling to the electorate.

Cross-platform botnet deployment – modern campaigns rarely limit themselves to a single platform. Using AI, malicious actors create interconnected bot networks that operate on multiple social media sites simultaneously. A coordinated approach might start a rumor on a fringe forum, amplify it via bots on Twitter, further propagate it through fake accounts on Facebook groups, and echo it in YouTube comments – all timed in concert. Such cross-platform coordination was observed in the earlier mentioned "Spamouflage" network linked to China, which spread content on Twitter, YouTube, Facebook, and even smaller platforms like Reddit and Tumblr, to maximize reach and create the appearance of a widespread movement. AI helps manage this complexity by centralizing control: operators can use dashboards to oversee their bot army's activities across different sites, ensuring the message is consistent and ubiquitous. This also makes disinformation more resilient; if one platform cracks down, the narrative persists elsewhere and can even be reintroduced to the original platform via users who encountered it on a different site.

Natural language generation in multiple languages and styles – large language models enable disinformation agents to automatically generate content that is linguistically and culturally tailored. GPT-4, for example, can write persuasive texts not just in English but in dozens of languages, capturing nuances that a non-native speaker (or older machine translation tools) might miss. This capability means a single campaign can target audiences in different countries with messages calibrated to local context. It also means that AI-written propaganda can mimic the style of particular communities (e.g., using youth slang on one forum, formal terminology on another) to blend in. NewsGuard, a firm that tracks online misinformation, reported in 2024 that some propaganda sites were using AI to write articles that adapt to the normative tone of particular regions, making the misinformation harder to flag by readers who expect certain idioms or references. This multilingual, multi-style generation greatly increases the reach of disinformation and makes global narratives possible – such as coordinated COVID-19 misinformation that simultaneously targeted audiences in the US, Europe, and Asia with different culturally specific angles.

The impact of this large-scale coordination is evident in several statistics from recent studies:

The sheer volume and reach of AI-coordinated content is staggering. By one estimate, toxic misinformation spread via coordinated networks reaches millions of users daily and has affected political processes in over 50 countries (Johnson et al., 2024). This global penetration suggests that no region is immune; wherever social media use is prevalent, so too is the potential for AI-boosted disinformation to interfere with public discourse.

The prevalence of bots as part of internet traffic highlights the magnitude of the challenge. As of 2024, bots accounted for an estimated 42% of all web traffic, and importantly, about 65% of those bots were classified as malicious (engaged in activities like scraping, spamming, and disinformation) (Akamai, 2024). Legitimate bots (like search engine crawlers) are outnumbered by those with nefarious purposes. This indicates that a significant portion of online interactions or "users" at any given time might actually be automated agents. The internet landscape is thus partly artificial, and public opinion metrics gleaned from online data need to be interpreted with caution, knowing a large fraction could be bot activity.

In social media specifically, infiltration by fake accounts is massive. On X/Twitter, researchers in 2022 identified roughly 16.5 million accounts that were dedicated solely to pushing false information or spam narratives (Akamai, 2024). These are accounts that virtually contribute nothing genuine to discourse, only propaganda or scams. It shows how platforms can be flooded with inauthentic entities, which can confuse users (especially new or less savvy ones who may not suspect that a personable profile with a smiling face is a bot in disguise).

Bots' ability to drive content is illustrated by specific events. During the first impeachment of U. S. President Donald Trump (late 2019 into 2020), it was found that bots generated 31% of the Twitter posts about the impeachment proceedings. These bots were often amplifying hyper-partisan or misleading narratives around the event (George Washington University, 2024). This level of automated involvement in a crucial national conversation raises questions about how public perception might have been swayed or distorted by inauthentic input.

Cross-platform connectivity – a study from George Washington University in 2024 revealed that disinformation operations routinely connect across an average of 23 different online platforms and services. This includes not just the big social media names, but also messaging apps, forums, comment sections of news sites, and more (George Washington University, 2024). By weaving threads through smaller niche sites (where moderation might be lax) into larger ones, campaigns can effectively launder the content – a false narrative might start in the dark corners of the web and then appear on Twitter cited as "seen on X forum," then get into mainstream Facebook posts, etc., chaining its way to respectability.

The presence of unreliable AI-generated news sites further complicates the landscape. NewsGuard identified over 1,100 websites in 2023 that were essentially content mills using AI to produce news-like articles, often riddled with inaccuracies or sensationalism (NewsGuard, 2024). These sites, operating in multiple languages, churn out a high volume of clickbait and propaganda, and they monetize through ads. Because they produce so much content, their stories frequently surface in search engine results or social media shares. If someone searches a topical question, they might end up on one of these sites without realizing its dubious nature. The proliferation of such sites means that traditional cues people used to judge credibility (professional layout, volume of content, etc.) are no longer reliable – AI can provide all those, yet the information can be false.

Deepfake proliferation ties into coordination as well. MarketsandMarkets research in 2024 indicated online deepfake videos were doubling every six months since 2019, reaching at least 95,820

by 2023. Many of these are pornographic or frivolous, but a portion are used in coordinated influence or harassment campaigns (MarketsandMarkets, 2024). The accelerating trend suggests that deepfakes may become a more regular feature in disinformation operations, as tools to generate them become user-friendly and accessible.

Economic and confidence impacts – we have mentioned the $78 billion economic cost of misinformation in 2020, a figure that likely includes various productivity losses, scams, and mitigative expenses (World Economic Forum, 2020). Additionally, public confidence measurements, like the Ipsos/UNESCO finding that over 60% believe AI makes creating fake news easier, reflect a growing awareness that technology is amplifying falsehoods, which might paradoxically lead to greater skepticism of legitimate news as well (once again feeding into the "nothing is true" post-truth problem).

Bringing together states, private firms, and other actors highlights the key players in this coordinated disinformation landscape:

State actors (like Russian intelligence, Chinese state-affiliated groups, and Iran's IRGC) use AI to scale up their information warfare and espionage activities. They typically have substantial resources and advanced tools (like Russia's Meliorator). They also target not only foreign populations but sometimes domestic audiences to shape internal narratives (U.S. Department of Justice, 2024a, 2024b).

Private "disinformation-for-hire" firms have emerged, offering their services to anyone willing to pay. These companies operate in the shadows, providing clients with bot networks, fake news site infrastructure, deepfake generation, and more. Their existence lowers the barrier for smaller regimes or even non-state groups (like extremist organizations or wealthy individuals) to run sophisticated influence campaigns without needing an in-house capability (Carnegie Endowment for International Peace, 2024).

Tech companies inadvertently are enablers when their AI innovations are repurposed. Companies like OpenAI, Google, and Meta are continuously improving AI's capabilities, which is beneficial for society in many ways, but also arms bad actors with better tools. There's a tension here: do these companies have a responsibility to restrict access to their models or build in disinfo safeguards? They have taken some steps (like OpenAI's content filters, or deepfake detection research by Facebook), yet the genie of general AI capability is out of the bottle and often open-source versions exist as well.

Media organizations and civil society are trying to respond by fact-checking, raising awareness, and creating initiatives like the Journalism Trust Initiative or deepfake databases. But they often play catch-up to the rapidly evolving tactics.

In face of this level of coordination, international cooperation and robust multi-stakeholder responses are critical. Already, collaborations like the NATO StratCom Center and the EU's East StratCom task force are sharing information on disinformation campaigns. Tech companies have formed alliances (e.g., the Global Internet Forum to Counter Terrorism, which could serve as a model for disinformation). But a more comprehensive global strategy may be required, as recommended earlier in the policy section: something akin to a standing coalition or information defense pact.

As AI continues to advance (with looming prospects of even more powerful generative models, multimodal systems that combine text, video, and other media, and possibly AIs autonomously strategizing influence operations), the coordination and scale of disinformation could increase further. However, that same AI can be harnessed on the defensive side. The next subsection examines how AI is used to detect and counter disinformation, which is the hopeful flip side of this challenging coin.

4.7 AI in counter-disinformation efforts

While AI has supercharged the scale and sophistication of disinformation campaigns, it is also a powerful force for defending the integrity of the information space. Governments, technology firms, and civil society are increasingly deploying AI-driven tools to detect, verify, and mitigate false content online. These systems are being used not only to trace nefarious actions but also to preempt emerging threats, creating a growing ecosystem of proactive digital defense. To ensure effective deployment, however, it is crucial to distinguish clearly between the underlying AI capabilities and the intent behind their use. The same technologies that generate synthetic content can, when responsibly applied, expose malicious campaigns, support verification, and empower users with reliable information.

4.7.1 Detection of inauthentic content and accounts

Machine learning models have become essential in identifying patterns characteristic of fake news, deepfakes, and coordinated disinformation activity. For example, Meta (formerly Facebook) uses AI-based systems to scan user content for known hoaxes or misleading claims, cross-referencing with fact-checking databases (Meta, 2023). Computer vision and audio analysis algorithms can detect manipulated videos or synthetic speech by examining inconsistencies in facial expressions or acoustic signatures. Initiatives such as the Deepfake Detection Challenge, organized in 2020, helped develop state-of-the-art classifiers that remain foundational in today's detection workflows (Dolhansky et al., 2020).

4.7.2 Network pattern analysis

Beyond content, AI is critical in identifying coordinated inauthentic behavior. Social media platforms apply graph-based algorithms to detect clusters of accounts that share anomalous characteristics – such as newly created profiles with AI-generated avatars engaging in synchronized posting. This methodology has proven effective in uncovering large-scale influence operations, as reported by Meta's Coordinated Inauthentic Behavior reports.

4.7.3 Automated fact-checking and stance detection

Natural Language Processing (NLP) enables the real-time extraction and assessment of claims within text. Tools such as Full Fact's AI platform scan news and live transcripts for fact-checkable statements, aligning them with verified information (Full Fact, 2022). Similarly, stance detection algorithms analyze a document's agreement or contradiction with known facts, flagging likely misinformation even before human review.

4.7.4 Image and metadata verification

Reverse image search tools like TinEye or Google Fact Check Explorer employ AI to uncover the original context of repurposed or deceptive images. More advanced models analyze visual

inconsistencies such as lighting mismatches, noise patterns, or shadow anomalies, allowing investigators to detect subtle manipulations in photographs shared across social platforms.

4.7.5 Content moderation at scale

Platforms increasingly rely on AI to enforce content moderation rules at a speed and volume that human teams cannot match. For instance, during the COVID-19 pandemic, Facebook used AI to detect and remove or label over 20 million pieces of misinformation in a single quarter (Facebook Transparency Report, 2021). These models operate based on pre-defined misinformation heuristics, keyword analysis, and community flagging behaviors.

4.7.6 Human-AI collaboration for monitoring

AI supports human analysts by filtering vast datasets to identify emerging narratives, disinformation trends, or suspicious user behavior in fringe forums and encrypted chats. This hybrid approach enables analysts to focus on high-impact threats, supported by AI-processed evidence of their reach, sentiment, and virality trajectories.

4.7.7 Attribution and forensics

AI-powered systems analyze behavioral, linguistic, and infrastructural signatures across campaigns to attribute them to coordinated actors. One notable case is the attribution of the "Ghostwriter" campaign targeting European audiences, where linguistic forensics and shared infrastructure revealed a common origin behind dispersed incidents (Ben-Natan et al., 2021). These findings aid enforcement actions, including sanctions or indictments.

4.7.8 User-facing verification tools

AI applications are also being developed for end-users in the form of browser extensions and mobile apps that offer real-time credibility assessments of online content. These systems might alert users to suspicious sources, highlight inconsistencies, or provide automated summaries with reliability indicators. Although still nascent, such tools mirror spam filters in email systems and can help cultivate critical media literacy at scale.

4.7.9 Training and resilience via simulation

Gamified applications like "Bad News" and "Harmony Square" use AI to simulate disinformation environments and help users build psychological immunity to manipulative narratives. These tools adapt to user behavior, using reinforcement learning to personalize challenges and enhance learning effectiveness (Roozenbeek and van der Linden, 2019).

4.7.10 Causal framework for AI misuse and defense

Across these examples, a clear causal pathway emerges: neutral technological capability (e.g., generative or predictive AI) → operationalization by malicious actors (e.g., coordinated inauthentic behavior) → measurable social harm (e.g., vaccine hesitancy, electoral interference). Importantly, each stage also provides a leverage point for ethical counter-action. The very tools exploited to scale disinformation – such as generative models or bot frameworks – can be repurposed for detection, simulation, and remediation. For instance, adversarial testing using synthetic bots helps platforms stress-test moderation filters, while generative models are used to train classifiers to recognize manipulated content.

5 Conclusion

In the evolving contest between disinformation and truth, AI is both a weapon and a shield. Its deployment must be grounded in an understanding that harm stems not from the existence of a tool, but from its application and the ecosystem of incentives surrounding it. Therefore, any policy framework must prioritize intent-based regulation, investment in cross-sector R&D, and transparency-by-design. As the information environment grows more complex, building a layered, adaptive, and evidence-based response that leverages the full spectrum of AI capabilities is the only viable path forward. This review underscores that safeguarding our digital public sphere is not a matter of banning technology, but of channeling it toward resilience, accountability, and collective intelligence.

Author contributions

AR: Conceptualization, Investigation, Writing – original draft. OM: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. VG: Conceptualization, 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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

Access Now. (2023). World Press Freedom Day: India's "fake news" law must not censor journalists. New York, NY, USA: Access Now.

ACI Worldwide. (2024). Deepfake fraud analysis and consumer impact report. Elkhorn, NE, USA: ACI Worldwide.

AI Marketing Advances. (2023). AI-based psychological targeting in digital marketing: Trends and risks. AI Marketing Advances [online resource].

Akamai (2024). Bot Traffic Report 2024. Cambridge, MA, USA: Akamai Technologies.

BBC News. (2023). False asylum seeker claims spread online before Stockport unrest. London, UK: BBC News.

Ben-Natan, O., Cohen, D., Cohen, I., Glazner, R., Zohar, D., Tuval, A., et al. (2021). Ghostwriter in the shell: expanding on Mandiant's attribution of UNC1151 to belarus. Somerville, MA, USA: Recorded Future (Insikt Group).

Boston Fed (2024). Synthetic identity fraud in the U.S. financial sector. Boston, MA, USA: Federal Reserve Bank of Boston.

BusinessDasher. (2024). Global Media Consumption Trends. BusinessDasher.

Carnegie Endowment for International Peace (2024). The global disinformation order: 2024 global inventory of organized social media manipulation. Washington, DC, USA: Carnegie Endowment for International Peace.

Centers for Disease Control and Prevention (2020–2021). COVID-19 vaccine misinformation monitoring system. Atlanta, GA: CDC.

Centers for Disease Control and Prevention (CDC). (2021). *COVID-19 Vaccine Myths and Misinformation Tracking Report U.S.* Atlanta, GA, USA: Department of Health and Human Services.

China Daily (2025). AI labeling to fight spread of fake info. Beijing, China: China Daily.

China Global Television Network (2020). Italian citizens thank China for COVID-19 aid. Beijing, China: CGTN.

DataReportal (2025a). Global Digital Overview: April 2025 Update: We Are Social and Meltwater.

DataReportal (2025b). Global Social Media Users 2025. London, UK & San Francisco, CA, USA: We Are Social & Meltwater.

Department of Homeland Security (2024). Russian Disinformation Campaign Disruption Report. Washington, DC: Department of Homeland Security.

Dergipark (2024). Fake News Detection Models and Social Media Analysis. Dergipark.

Disinformation Research. (2023). Global response times to viral disinformation: A 3-year benchmark. *Disinformation Research*.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Bitton, J., Fang, S., et al. (2020). The Deepfake Detection Challenge. (DFDC) Preview Dataset. arXiv [Preprint]. 1–10. doi: 10.48550/arXiv.2006.07397.

European Parliament (2022). Deepfakes: MEPs warn of threats to democracy. Brussels: European Parliament.

Exploding Topics. (2025). Social Media Usage Statistics. Boston, MA, USA: Exploding Topics.

Election Commission of India. (2024). Voter Awareness and Disinformation Mitigation Measures: 2024 General Election Briefing. New Delhi, India: Election Commission of India.

Facebook Transparency Report (2021). Community Standards Enforcement. Menlo Park, CA, USA: Facebook.

Freedom House. (2023). Brazil's Social Media Regulation and the Fight Against Disinformation. Washington, DC, USA: Freedom House.

Full Fact. (2022). Automated Fact-Checking Tools and Platforms. London, UK: Full Fact.

Full Fact. (2023). No evidence supports viral asylum seeker violence claim in Stockport. London, UK: Full Fact.

George Washington University (2024). Computational Propaganda Research Project: 2024 Global Inventory of Social Media Manipulation. Washington, DC: George Washington University.

Grand View Research (2024). AI Text Generator Market Size, Share & Growth Report, 2030. San Francisco, CA: Grand View Research.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science. *Advances* 5:eaau4586. doi: 10.1126/sciadv.aau4586

Hillmann, P. (2023). The Deepfake That Fooled Crypto Executives. Cayman Islands: Binance Blog.

Hunter, C. (2024). My Experience as a Victim of Deepfake Pornography: Personal Blog. [online].

Hunter, J. (2024). Crisis amplification and the algorithmic public sphere: Lessons from real-time misinformation incidents. *The Journal of Crisis Media*. [Ahead of print].

Imperva/Statista. (2024). Share of bot traffic on the internet worldwide from 2016 to 2023. Statista, Hamburg, Germany.

Ipsos and UNESCO (2024). Global Survey on Internet Users' Experience with Disinformation. Paris: UNESCO.

Johnson, N. F., Sear, R., and Illari, L. (2024). Controlling bad-actor-artificial intelligence activity at scale across online battlefields. *PNAS Nexus*, 3, pgae004. doi: 10.1093/pnasnexus/pgae004

Kont, K., Kalmus, V., Siibak, A., and Masso, A. (2024). Children's strategies for recognizing and coping with online disinformation: A European comparison. *Safety*. 10:10. doi: 10.3390/safety10010010

MarketsandMarkets. (2024). Deepfake AI Market—Global Forecast to 2030. Pune, India: MarketsandMarkets.

Mehrotra, A., and Upadhyay, A. (2025). India's growing misinformation crisis: a threat to democracy. Washington, DC, USA: The Diplomat.

Meta (2023). Coordinated Inauthentic Behavior Report: Q2 2023. CA: Menlo Park.

Moore, M. (2024). Top cybersecurity threats to watch in 2025. San Diego, CA, USA: University of San Diego.

NewsGuard (2024). Tracking AI-Generated Misinformation: A 2024 Report. New York: NewsGuard.

Norton, R., and Marchal, N. (2023). Synthetic Amplification and Local Disorder: Case Studies in Digital Propaganda. Oxford, UK & New York, NY, USA: Oxford Internet Institute & Graphika.

Ofcom (2023). Online Nation 2023 Report. London: Ofcom.

Onfido (2023). Identity Fraud Report 2023. London: Onfido.

Parker, L., Byrne, J. A., Goldwater, M., and Enfield, N. (2021). Misinformation: an empirical study with scientists and communicators during the COVID-19 pandemic. *BMJ Open Sci* 5:e100188. doi: 10.1136/bmjos-2021-100188

Paul, K. (2023). Brazil receives pushback from tech companies on 'fake news' bill. London, UK: The Guardian.

Pennycook, G., McPhetres, J., Bago, B., and Rand, D. G. (2021). Attitudes about COVID-19 in Canada, the United Kingdom, and the United States: A comparative study. *Psychol. Sci.* 32, 2023–2036. doi: 10.1177/0956797620939054

PNAS Nexus. (2024). AI bots and disinformation: Network dynamics and intervention strategies. PNAS Nexus, *3*, 112–126.

Redline Digital. (2023). Global disinformation analytics: Fraud and microtargeting trends. Vilnius, Lithuania: Redline Digital.

Rossetti, M., and Zaman, T. (2023). Bots, disinformation, and the first impeachment of U.S. President Donald Trump. *PLOS ONE*, 18, e0283971. doi: 10.1371/journal.pone.0283971

Roozenbeek, J., and van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *J. Risk Res.* 22, 570–580. doi: 10.1080/13669877.2018.1443491

Security.org. (2024). State of deepfakes 2024: Proliferation and public awareness. Security.org, Los Angeles, CA, USA.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nat. Commun.* 9:4787. doi: 10.1038/s41467-018-06930-7

Slovak Ministry of Interior Ministry of the Interior of the Slovak Republic (2024). Report on Election Integrity and Disinformation. Bratislava.

SOAX. (2025). Average Time Spent on Social Media in 2025. SOAX, London, UK.

Stanford Internet Observatory (2024). Spamouflage Unmasked: An Analysis of a Pro-PRC Influence Operation. Stanford, CA: Stanford Internet Observatory (Stanford University).

Sumsub. (2024). Identity verification and synthetic fraud report 2023. Limassol, Cyprus: Sumsub.

U.S. Department of Justice (2024a). Indictment of Iranian Nationals for Cyber-Enabled Disinformation and Threat Campaign. Washington, DC: U.S. Department of Justice.

U.S. Department of Justice (2024b). Joint Statement on Disruption of Russian AI-Enhanced Disinformation Network. Washington, DC: U.S. Department of Justice.

U.S. Department of State (2024). Countering State-Sponsored Disinformation 2024 Report. Washington, DC: U.S. Department of State.

UK Electoral Commission (2016). Report on the 23 June 2016 referendum on the UK's membership of the European Union. London: UK Electoral Commission.

University of Southern California (2020). Coordinated amplification network analysis: COVID-19 reopen America campaign. Los Angeles: USC.

Van der Linden, S., Roozenbeek, J., and Compton, J. (2021). Inoculating against fake news about COVID-19. *Front. Psychol.* 11:566790. doi: 10.3389/fpsyg.2020.566790

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

World Economic Forum (2020). The Global Risks Report 2020. *World Economic Forum*. Geneva: World Economic Forum.

World Health Organization (2018). Artificial Intelligence for Health: World Health Assembly 71. Geneva: WHO.

Zhang, Y., Song, W., Koura, Y. H., and Su, Y. (2023). Social Bots and Information Propagation in Social Networks: Simulating Cooperative and Competitive Interaction Dynamics. *Systems* 11:210 doi: 10.3390/systems11040210

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 11 September 2025
DOI 10.3389/frai.2025.1566022

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
Abel Suing,
Universidad Técnica Particular de Loja, Ecuador
Arif Zainudin,
Universitas Pancasakti Tegal, Indonesia

*CORRESPONDENCE
Andrii Paziuk
✉ andrii.paziuk@npp.kai.edu.ua

RECEIVED 24 January 2025
ACCEPTED 18 August 2025
PUBLISHED 11 September 2025

CITATION
Paziuk A, Lande D, Shnurko-Tabakova E and Kingston P (2025) Decoding manipulative narratives in cognitive warfare: a case study of the Russia-Ukraine conflict.
Front. Artif. Intell. 8:1566022.
doi: 10.3389/frai.2025.1566022

Decoding manipulative narratives in cognitive warfare: a case study of the Russia-Ukraine conflict

Andrii Paziuk[1]*, Dmytro Lande[2], Elina Shnurko-Tabakova[3] and Phillip Kingston[1]

[1]Law and International Relations Faculty, State University Kyiv Aviation Institute, Kyiv, Ukraine, [2]Department of Information Security, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, [3]International Institute of Cyber Diplomacy and AI Security, Law and International Relations Faculty, State University Kyiv Aviation Institute, Kyiv, Ukraine

Introduction: This study investigates the construction and dissemination of manipulative narratives in the context of cognitive warfare during the Russia-Ukraine conflict. Leveraging a mixed-methods approach that integrates AI-assisted semantic analysis with expert validation, we examine how adversarial messaging exploits cognitive biases-such as fear and confirmation bias-to influence perceptions and disrupt institutional trust.

Methods: Using the proprietary Attack-Index tool and large language models (LLMs), we detect linguistic markers of manipulation, including euphemisms, sarcasm, and strategic framing.

Results: Our findings demonstrate that emotionally charged narratives, particularly those invoking nuclear threat scenarios, are synchronized with key geopolitical events to influence decision-makers and public opinion. The study identifies five thematic clusters and traces shifts in rhetorical strategies over time, showing how manipulative discourse adapts to geopolitical contexts. Special attention is given to the differentiated targeting of international political elites, Western publics, and Russian domestic audiences, each exhibiting varied cognitive vulnerabilities.

Discussion: We acknowledge methodological and ethical limitations, including the dual-use potential of AI tools and challenges in establishing causal inferences. Nonetheless, this study offers the following key contributions:

1. Empirically establishing nuclear rhetoric as a strategic element of narrative manipulation, particularly around NATO summits and military aid announcements.
2. Advancing an integrated analytical framework that combines semantic clustering and AI-based discourse detection to monitor information threats in real time.
3. Providing actionable insights for policy and digital security, including the development of countermeasures and international collaboration in addressing cognitive warfare.

KEYWORDS

cognitive warfare, semantic analysis, Russia-Ukraine conflict, nuclear threat narratives, emotional manipulation, disinformation

1 Introduction

Cognitive warfare represents a new frontier in the evolution of conflict, where the mind itself becomes the primary battleground. Unlike traditional kinetic warfare, cognitive warfare operates in the psychological and informational domains, exploiting vulnerabilities in human cognition to manipulate beliefs, emotions, and decision-making processes.

NATO describes it as a deliberate effort to influence, disrupt, or protect the cognitive processes of individuals and societies, highlighting its growing role in modern hybrid threats (Deppe and Schaal, 2024). This emergent form of warfare transcends conventional propaganda and psychological operations by incorporating advanced technologies and strategic timing to maximize its impact. Cognitive warfare has been recognized as a critical element of hybrid threats, which combine conventional, unconventional, and technological methods to achieve strategic objectives. Its operational methods include the dissemination of disinformation, narrative framing, emotional manipulation, and exploiting cognitive biases such as confirmation bias and anchoring. The strategic use of these techniques is evident in modern conflicts, where adversaries target individuals and societal structures to destabilize trust in institutions and polarize populations. NATO's reports emphasize the role of cognitive warfare in shaping perceptions during geopolitical conflicts, highlighting its integration with cyber operations and kinetic attacks (NATO Allied Command Transformation, 2023). This multifaceted approach leverages digital ecosystems to create information asymmetry, where adversaries overwhelm their targets with a flood of narratives, some true, some false, and some a blend of both. By doing so, they obscure the truth, manipulate public opinion, and erode trust in democratic institutions. This tactic has been prominently observed in the Russia-Ukraine conflict, where disinformation campaigns have been used to undermine support for Ukraine and amplify divisions within NATO and the European Union (Digital Forensic Research Lab, 2023). Technological advancements, particularly in artificial intelligence, big data analytics, and social media algorithms, have amplified the reach and precision of cognitive warfare. Platforms like X (formerly Twitter) and Facebook have become battlegrounds for influencing public discourse, with adversaries employing bot networks, troll farms, and deepfake technologies to propagate narratives. These operations are reactive and proactive, often synchronizing their efforts with significant geopolitical events such as elections, military exercises, or international summits (Marsili, 2023). The ability to control the narrative in these moments can significantly influence electoral outcomes, consultative democratic processes, direct democracy, policy decisions and support, alliances, traditional media, and overall public support.

1.1 Significance of the research

The digital age has exponentially increased the scope and impact of cognitive warfare, making it a critical area of study for ensuring geopolitical stability. As societies become more interconnected through digital platforms, the opportunities for adversaries to exploit cognitive vulnerabilities have grown. Cognitive warfare poses unique challenges because its effects are not always immediately visible. Unlike physical attacks, the damage inflicted by cognitive operations is often psychological and societal, manifesting over time as declining trust, increasing polarization, and eroding democratic norms. The importance of understanding cognitive warfare lies in its capacity to disrupt the foundational elements of liberal democracies. Cognitive warfare undermines the social contract that binds societies together by targeting trust. In a world where information is currency, the ability to control or manipulate narratives becomes a powerful weapon. This has been demonstrated in conflicts such as the Russia-Ukraine war, where disinformation campaigns have aimed to delegitimize Ukrainian leadership, weaken international support for Ukraine, and sow discord among NATO allies. Examples such as claims of Ukraine planning to use a "dirty bomb" to frame Russia illustrate the tactical deployment of fear-inducing narratives designed to polarize international opinion. Moreover, cognitive warfare extends beyond the battlefield, affecting policymaking, election outcomes, and international relations. Governments and organizations must recognize that cognitive warfare is not just a military issue but a societal one, requiring a whole-of-society response. Media literacy programs, public awareness campaigns, and collaboration between governments, academia, and the private sector are essential to building resilience against cognitive attacks. This research significantly advances the understanding of cognitive warfare, particularly its application during the Russia-Ukraine conflict. It identifies nuclear rhetoric as a pivotal strategic tool within manipulative narratives, demonstrating their synchronization with geopolitical milestones such as NATO summits and military aid announcements. By introducing an innovative analytical framework, which combines semantic analysis and AI-driven methodologies, the study enhances the detection and prediction of disinformation narratives, providing actionable insights for countering their psychological and geopolitical impacts. Central to this research is the concept of leveraging predictive analytics in cognitive warfare. Tools like the Attack Index and advanced AI-driven methodologies are employed to forecast manipulative narrative trends, bridging the gap between retrospective analysis and proactive strategic planning. This approach not only reveals the mechanisms of cognitive operations but also underscores the necessity of aligning counter-narratives and policy responses preemptively. Through the lens of the Russia-Ukraine conflict, the study illustrates how cognitive warfare strategically employs fear and uncertainty, particularly through nuclear rhetoric, to manipulate public discourse and achieve geopolitical objectives. This synchronization of manipulative narratives with critical geopolitical events exemplifies the precision and adaptability of modern cognitive warfare tactics.

1.2 Research gap

While the study of cognitive warfare has gained traction in recent years, significant gaps remain in the literature. Much of the existing research focuses on the tactical aspects of disinformation, such as the dissemination methods and the role of social media platforms. However, there is a need for a more comprehensive understanding of the strategic dimensions of cognitive warfare, particularly its integration with other elements of hybrid threats, such as cyber operations and economic coercion. One underexplored area is the synchronization of cognitive operations with geopolitical milestones. For example, during the Russia-Ukraine conflict, disinformation campaigns have been carefully timed to coincide with NATO summits, military aid

announcements, and key elections in Western democracies. This strategic timing magnifies the impact of cognitive operations by aligning them with moments of heightened public and political attention. Understanding how adversaries coordinate these efforts is crucial for developing effective countermeasures (Marsili, 2023).

Another gap lies in the study of emotional manipulation in cognitive warfare. While much attention has been given to the content of disinformation, less has been said about its emotional appeal. Adversaries often craft narratives that evoke fear, anger, or resentment, knowing that emotions significantly shape beliefs and behaviors. For instance, during the Russia-Ukraine war, disinformation campaigns have exploited fears of nuclear escalation and economic instability to influence public opinion in Europe and the United States (Splidsboel Hansen, 2021). Researching the emotional dimensions of cognitive warfare can provide deeper insights into its effectiveness and inform the development of more nuanced counter-narratives. Finally, there is a need to explore the role of emerging technologies in detecting and countering cognitive warfare. Semantic analysis, sentiment detection, and narrative mapping have shown promise in identifying disinformation patterns. However, these tools must be integrated into a broader framework that accounts for cognitive operations' strategic and emotional dimensions. Developing such a framework is essential for staying ahead in this evolving domain of conflict.

1.3 Research objectives

This research aims to address these gaps by examining the mechanisms, impacts, and countermeasures of cognitive warfare, focusing on the Russia-Ukraine conflict. The specific objectives are as follows:

1. Analyze the mechanisms of cognitive warfare: investigate how adversaries use narrative framing, emotional manipulation, and cognitive biases to influence public opinion and decision-making.
2. Explore the strategic integration of cognitive operations: examine how cognitive warfare is synchronized with other elements of hybrid threats, such as cyberattacks and economic coercion, to maximize its impact.
3. Evaluate the role of timing and context: study the timing of cognitive operations about geopolitical events, identifying patterns and strategies used by adversaries.
4. Propose countermeasures and resilience-building strategies: develop recommendations for governments, organizations, and societies to enhance their resilience against cognitive warfare, focusing on public awareness, media literacy, and advanced analytical tools.

2 Literature review

2.1 Foundational theories of cognitive warfare

Cognitive warfare represents a pivotal evolution in strategic conflict, blurring the lines between psychological operations, information warfare, and hybrid threats. Rooted in psychological theories of influence and propaganda, cognitive warfare leverages advances in digital technology to target the cognitive vulnerabilities of individuals and societies (Bilal, 2021). This framework has been extensively discussed by Deppe and Schaal (2024), who describe cognitive warfare as an "existential threat to the cognitive domain," where perception becomes the primary battleground. Building on foundational theories, NATO's reports outline how cognitive warfare integrates with hybrid strategies, targeting systemic and individual-level cognition to create asymmetries in trust, perception, and decision-making (NATO Allied Command Transformation, 2024). The hybrid warfare lens, as outlined by Weissmann et al. (2021), situates cognitive warfare within a broader spectrum of unconventional conflict, merging cyber, kinetic, and cognitive strategies. Historical antecedents of cognitive warfare can be traced to Cold War psychological operations, where state actors employed propaganda to influence ideological allegiance. Buzan and Waever (2003) theories on securitization add context to the weaponization of narratives, framing them as tools to mobilize collective fears and justify political actions. These theoretical underpinnings have gained renewed relevance in the digital era, where social media and artificial intelligence amplify the reach and effectiveness of cognitive warfare (Ngwainmbi, 2024).

2.2 Mechanisms of cognitive warfare

2.2.1 Psychological manipulation

Psychological manipulation remains a cornerstone of cognitive warfare, targeting emotional and cognitive vulnerabilities to induce behavioral and perceptual shifts. Fear, anger, and existential threats are central to this strategy. Henschke (2025) argues that psychological manipulation in cognitive warfare often operates at a subconscious level, leveraging emotional salience to bypass rational scrutiny. In the context of the Russia-Ukraine conflict, psychological manipulation has been used to frame existential threats, such as NATO's alleged encroachment on Russian sovereignty, as a means to justify aggressive actions. Gambhir (2016) underscores the parallels between these tactics and ISIS's use of fear-based narratives, which targeted vulnerabilities in Western societies to radicalize individuals and undermine public confidence.

2.2.2 Narrative synchronization

Narrative synchronization refers to aligning manipulative narratives with key geopolitical events, enhancing their resonance and perceived legitimacy. Jensen and Ramjee (2023) highlight how Russian narratives during the Ukraine crisis synchronized with NATO summits and Western military aid announcements, framing these actions as provocations. This synchronization creates temporal relevance, which amplifies the perceived credibility of disinformation. Marsili (2023) describes how narrative synchronization is facilitated through algorithmic targeting, ensuring that key messages reach specific audiences at opportune moments. By aligning narratives with tangible events, such as energy shortages or sanctions, adversaries create a feedback loop that reinforces public skepticism and divisions.

2.2.3 Exploitation of cognitive biases

Exploiting cognitive biases such as confirmation bias, anchoring, and the availability heuristic is integral to cognitive warfare strategies. Nguyen (2023) explains that these biases predispose individuals to accept disinformation that aligns with their pre-existing beliefs, making them more susceptible to manipulation. Extensive research, including Benkler et al. (2018) and Bradshaw and Howard (2019), demonstrates how state-sponsored disinformation exploits cognitive biases to polarize audiences and erode trust in democratic institutions. Plaza et al. (2023) identify emotional contagion as a related mechanism, where emotionally charged narratives spread rapidly through social networks, amplifying their impact. This dynamic highlights the interplay between psychological vulnerabilities and digital platforms in cognitive warfare. Lewandowsky et al. (2017) highlight mechanisms that make disinformation effective and argue that simple, emotionally resonant narratives are more easily processed and remembered by individuals than complex, nuanced truths. This cognitive preference for simplicity makes populations more susceptible to accepting and spreading disinformation, especially when the information is intricate and demands more significant cognitive effort to understand.

2.3 Role of AI in amplifying cognitive warfare

2.3.1 AI-driven content creation

Artificial intelligence has revolutionized cognitive warfare by enabling the creation of highly convincing and scalable manipulative content. Deepfakes, synthetic media, and automated narrative generation tools allow adversaries to craft persuasive disinformation with minimal effort (Henschke, 2025). For example, AI-generated deepfake videos of Ukrainian leaders were used to undermine public confidence and create confusion during the early stages of the conflict (Fisher et al., 2024).

2.3.2 Sentiment analysis and targeted messaging

AI tools like sentiment analysis play a dual role in cognitive warfare. On one hand, they enable adversaries to monitor public sentiment and tailor narratives to resonate with emotional undercurrents. Huang et al. (2024) describe how sentiment analysis algorithms were used to detect and exploit shifts in public opinion during geopolitical crises. On the other hand, these tools empower counter-cognitive operations, enabling real-time detection of disinformation trends (Index Systems Ltd., 2023). Modern AI has enabled a deeper understanding of the propagation of messaging through the social graph that underpins social networks. This enriches the ability to granularly segment users and their patterns of influence, and personalize disinformation to that segment and its most effective delivery mechanism.

2.3.3 Strategic amplification of narratives

The scalability of AI-driven tools amplifies the reach and impact of disinformation campaigns. Nimmo and Flossman (2024) highlights the role of large-scale language models in generating tailored narratives that exploit cultural and ideological divisions. This strategic deployment creates a multiplier effect, where the same narrative resonates across diverse audiences, maximizing its disruptive potential.

2.4 Countermeasures against cognitive warfare

2.4.1 Technological interventions

Technological solutions form the backbone of counter-cognitive warfare strategies. Attack-Index allows real-time monitoring of adversarial narratives, enabling rapid response to disinformation campaigns. Additionally, AI-driven detection systems identify linguistic patterns and anomalies that indicate manipulative intent (Huang et al., 2024).

2.4.2 Media literacy and public awareness

Building societal resilience through media literacy programs is essential to countering cognitive warfare. NATO Allied Command Transformation (2023) emphasizes the role of public education initiatives in equipping individuals with the critical thinking skills needed to identify and resist disinformation. Sarwono (2022) underscores the importance of targeting these programs at vulnerable demographics, such as youth and digitally marginalized populations, who are disproportionately affected by manipulative narratives.

2.4.3 Policy and collaborative frameworks

International collaboration is critical in addressing the transnational nature of cognitive warfare. Orinx and Struye de Swielande (2022) call for harmonized policy frameworks that standardize responses to cognitive threats across allied nations. These frameworks enhance collective resilience and serve as deterrents by increasing the costs of engaging in cognitive warfare. The literature on cognitive warfare reveals its central role in shaping modern geopolitical conflicts, emphasizing its evolution from traditional propaganda to highly sophisticated digital strategies. Foundational works such as NATO's Cognitive Warfare Framework and other key contributions (e.g., NATO Allied Command Transformation, 2024; Splidsboel Hansen, 2021) contextualize cognitive warfare as a method for undermining trust, polarizing societies, and destabilizing political structures. Key mechanisms of cognitive warfare—psychological manipulation, narrative synchronization, and emotional exploitation—are strategically employed to amplify societal vulnerabilities. Studies such as Huang et al. (2024) and Grindrod (2024) demonstrate how emotional exploitation targets individual biases, including confirmation bias and availability heuristics, leveraging these tendencies to promote divisive narratives. By aligning narratives with pivotal geopolitical events, adversaries create a feedback loop that strengthens disinformation and delays collective responses (Brusylovska and Maksymenko, 2023). The integration of advanced technologies further elevates the efficacy of cognitive

warfare. Works by Nguyen (2023) and Nimmo and Flossman (2024) illustrate how tools like deepfakes, sentiment analysis, and algorithm-driven narratives enhance the precision of disinformation campaigns. These tools are instrumental in geopolitical conflicts, such as the Russia-Ukraine war, where adversaries use AI-driven methods to erode public trust and manipulate global perceptions (Benkő and Biczók, 2024).

Current countermeasures reflect the complexity of combating cognitive warfare. Media literacy programs (e.g., Wallenius, 2023), real-time monitoring tools like Attack-Index (Index Systems Ltd., 2022, 2023), and AI-driven detection systems are the solutions proposed to mitigate these threats. However, as the review highlights, the adaptability of adversaries and the limitations of current frameworks necessitate ongoing innovation and collaboration across sectors (Jensen and Ramjee, 2023). Benkler et al. (2018) analyze the structure and dynamics of media ecosystems and argue that the commitment to free speech, especially in digital and social media platforms, can inadvertently provide a fertile ground for disinformation campaigns. Furthermore, Bradshaw and Howard (2019) demonstrate how disinformation campaigns evade regulatory attention and enforcement action by masquerading as legitimate political discourse. However, attempts by governments or social media platforms to intervene can paradoxically validate these falsehoods, as regulatory measures are sometimes perceived as acknowledgments of the underlying credibility of the suppressed narratives (DiResta, 2020; Pennycook et al., 2020). Such interventions may reinforce conspiratorial beliefs, leading to increased trust in disinformation among specific populations (Tucker et al., 2017; Marwick and Lewis, 2017). In sum, the literature illustrates the multidimensional nature of cognitive warfare, which intersects psychology, technology, and geopolitics. While advancements in AI and digital platforms have enhanced the reach and impact of manipulative campaigns, they also provide opportunities for countering these threats through detection and resilience-building measures. Future research must focus on bridging existing gaps, particularly integrating interdisciplinary approaches to address the evolving tactics of cognitive warfare. This review underscores the need for a proactive and collaborative global response to safeguard societal stability in the face of this emerging domain.

3 Methodology

3.1 Research design

This study employs a comprehensive mixed-method approach integrating qualitative and quantitative methodologies to analyze manipulative narratives in cognitive warfare. By combining semantic analysis through the Attack-Index tool with AI-driven narrative analysis, the methodology captures the intricacies of narrative construction, emotional manipulation, and the psychological impact on target audiences. This design bridges theoretical gaps in understanding cognitive warfare and its practical manifestations, providing a holistic lens for examining the intersection of manipulative strategies and technology (Plaza et al., 2023; Deppe and Schaal, 2024; Henschke, 2025).

3.2 Rationale for a mixed-method approach

Cognitive warfare involves the interplay of qualitative dimensions—like understanding emotional and rhetorical subtleties—and quantitative dimensions—such as the prevalence, intensity, and temporal dynamics of narratives. The mixed-method approach ensures both depth and breadth, offering nuanced insights into how manipulative narratives evolve and resonate with target audiences (NATO Allied Command Transformation, 2023; Weissmann et al., 2021).

3.3 Key components of the research design

3.3.1 Semantic analysis through the attack-index tool

The Attack-Index tool serves as a comprehensive framework for analyzing manipulative narratives by identifying recurring themes, emotional triggers, and narrative trajectories. Its quantitative capabilities measure the frequency and intensity of specific themes, such as nuclear threats, across extensive datasets, offering precise numerical insights into the scale and prevalence of disinformation. Simultaneously, its qualitative features delve into linguistic framing and emotional undertones, revealing the psychological tactics employed in narrative construction (Index Systems Ltd., 2023). The Attack-Index tool was pivotal in uncovering key phraseological patterns and narrative shifts, such as the recurring theme of nuclear aggression, quantified by the tool's clustering algorithm (Appendix). This capability enables the identification of key narrative lines and their dynamics, offering a deeper understanding of how manipulative narratives evolve over time. By clustering related linguistic elements, the tool highlights the structural coherence of disinformation campaigns, aiding in pinpointing central themes and their trajectories (Lande and Shnurko-Tabakova, 2019). As an essential instrument in countering cognitive warfare, the Attack-Index enables the prompt identification of disinformation campaigns, thereby supporting the development of targeted countermeasures. Through its capacity for real-time monitoring and retrospective analysis, it not only helps detect emerging threats but also informs strategic decision-making. The integration of the Attack-Index into broader cognitive warfare frameworks strengthens efforts to neutralize manipulative narratives, safeguard the information environment, and build resilience against information threats (Lande, 2024).

3.4 AI-driven narrative analysis

Advanced AI tools, including pre-trained models and Natural Language Processing (NLP) algorithms, supplement semantic analysis by identifying subtler manipulations such as sarcasm, euphemisms, and framing techniques. These tools enhance the detection of cognitive biases and rhetorical tactics embedded in disinformation (Huang et al., 2024; Grindrod, 2024).

3.5 Temporal and event-based analysis

By correlating shifts in narratives with geopolitical events, such as NATO summits or military aid announcements, this component provides insights into how manipulative narratives are synchronized with broader strategies to maximize psychological impact (Monaghan, 2020; Benkő and Biczók, 2024).

3.6 Data collection: a comprehensive framework

Effective data collection is central to analyzing manipulative narratives in cognitive warfare. This study adopts a systematic, multi-source data collection strategy to ensure the scope and depth necessary for meaningful analysis.

3.7 Data sources

The system analyzed around 4,000 Russian websites and 3,000 Telegram channels from November 1, 2022, to March 5, 2023. Using linguistic analysis and machine learning, 36,821 phrases were identified, of which 170 were selected for further analysis. Of these, 170 key phrases—such as "nuclear attack," "dirty bomb," "Kyiv authorities," and "nuclear treaties"—were selected for further analysis. These persistent phrases formed the foundation for constructing a narrative network and uncovering the structural dynamics of disinformation campaigns. Disinformation narratives were categorized into five key clusters, including "Kyiv authorities and the dirty bomb" and "Special operation and nuclear energy," as visualized in the cluster network map in Appendix Figure 5. This clustering highlights the diversity and strategic focus of Russian narratives.

3.8 Preprocessing techniques

Tokenization: text data were parsed into tokens to facilitate semantic and linguistic analysis. Noise filtering: irrelevant or redundant content was removed to ensure analytical rigor. Semantic tagging: data were annotated to categorize emotional tones, themes, and rhetorical devices. Temporal segmentation: data were segmented to analyze narrative evolution over time and correlate shifts with geopolitical events.

3.9 Definition of tokens

Let the set of tokens in the network be defined as:

$$T = \{t_1, t_2, \ldots, t_n\}$$

where t_i represents either an individual word in its normalized form (e.g., lemmatized word) or a stable phrase (multi-word expression).

3.10 Network representation

The network of phrase interconnections is represented as a graph:

$$G = (V, E)$$

where:

- V is the set of vertices, corresponding to tokens T.
- $E \subseteq V \times V$ is the set of edges, representing relationships or co-occurrences between tokens.

3.11 Edge weights

Each edge $e_{ij} \in E$ has a weight w_{ij}, representing the strength of the connection between tokens t_i and t_j. The weight can be calculated based on metrics such as frequency of co-occurrence in the text corpus:

$$w_{ij} = f(t_i, t_j)$$

where $f(t_i, t_j)$ is the frequency of co-occurrence of t_i and t_j within a defined window size or context.

3.12 Cluster analysis

The Gephi environment was employed to analyze narrative data by clustering tokens into modularity classes, thereby identifying distinct themes or "concept classes" within the dataset. The use of network visualizations, such as those generated with Gephi, provided critical insights into narrative interconnections and thematic clustering, revealing how narratives evolved in response to geopolitical developments (Appendix Figure 5). This approach is grounded in graph theory and modularity optimization, enabling the detection of cohesive narrative clusters that reveal the structural dynamics of disinformation campaigns (Bruns and Snee, 2022). Modularity optimization is central to cluster detection. It quantifies the strength of division of a network into clusters (or communities). In this study, various modularity methods were considered (Traag et al., 2019), and the Potts model (Wu, 1982) was specifically applied. The Potts model incorporates a resolution parameter γ, which adjusts the granularity of clustering, allowing the system to determine the optimal number of clusters.

Quality function $H(G, P)$, or briefly $H(P)$, for the partition P into modules (clusters) of the graph G is written as:

$$H(P) = -\sum_{C \in P} e_C - \gamma \cdot n_C^2,$$

where each cluster $C \in P$ consists of e_C edges and n_C nodes, and γ is the resolution parameter, which significantly influences the partitioning of the graph into clusters.

3.13 Clustering algorithm

The clustering procedure employed within the Gephi environment followed a structured algorithm designed to identify and classify thematic concentrations within the dataset. The integration of modularity clustering and token detection, as detailed in the Appendix, provided a comprehensive framework for analyzing narrative cohesion and thematic evolution across disinformation campaigns. This approach ensured methodological rigor and reproducibility in analyzing the Russian disinformation narratives.

3.13.1 Initialization

Nodes, representing individual tokens or phrases, were preliminarily distributed into clusters based on their initial associations. This step established a foundational configuration for subsequent modularity optimization.

3.13.2 Evaluation

The modularity of the current cluster configuration, denoted as H(P), was calculated. Modularity quantifies the quality of the network partitioning by measuring the strength of connections within clusters relative to those between clusters.

3.13.3 Merging clusters

Nodes or groups of nodes were iteratively merged to enhance modularity. This step optimized the grouping of nodes by forming clusters with strong internal connections.

3.13.4 Iteration

The iterative process of merging and evaluation continued until maximum modularity was achieved, indicating the optimal partitioning of the network. Once modularity ceased to improve, the process was finalized.

3.13.5 Finalization

The algorithm produced well-defined modularity classes, or clusters, characterized by high internal cohesion and minimal external overlap. These clusters encapsulated distinct thematic narratives within the dataset.

3.14 Semantic analysis using attack-index

The Attack-Index tool is central to this study's methodology, offering a sophisticated approach to uncovering emotional triggers, key themes, and narrative shifts over time. Designed to track linguistic patterns and emotional cues, the tool provides insights into constructing and disseminating manipulative narratives. Its utility in identifying subtle yet impactful rhetorical devices—such as geopolitical framing and existential threats—is particularly relevant in cognitive warfare. The semantic algorithms employed by the Attack-Index are tailored to detect nuanced narrative shifts, particularly during critical geopolitical events. The procedure involved the following steps:

3.14.1 Identifying recurring motifs

The tool flagged recurring themes such as existential risks, nuclear threats and perceived aggression from Western alliances. These motifs were prevalent in Russian state-controlled media, aligning with broader strategies to provoke fear and justify aggressive policies.

3.14.2 Tracking emotional tone over time

Changes in emotional tone were linked to specific geopolitical events, such as NATO summits or military mobilizations.

3.14.3 Linking themes to events

Semantic tagging connected themes to specific events, such as Western sanctions or high-profile diplomatic meetings. These connections revealed how disinformation was synchronized with real-world developments to amplify its psychological impact.

3.15 AI-driven analysis

To complement the insights from the Attack-Index, the study employed AI-driven tools, including pre-trained Large Language Models (LLMs), to analyze the subtleties of manipulative narratives. Unlike traditional semantic tools, LLMs excel at identifying contextual manipulations, such as sarcasm, euphemisms, and logical inconsistencies. These capabilities are crucial for understanding how disinformation exploits psychological vulnerabilities, including biases like confirmation bias and availability heuristics (Grindrod, 2024; Huang et al., 2024).

3.16 Detecting linguistic patterns

At the core of the analysis was the deployment of pre-trained AI models designed to detect subtle manipulations embedded within language. These models identify framing techniques that are not immediately apparent to human analysis. For instance, narratives were often framed to downplay acts of aggression while simultaneously amplifying victimhood to evoke empathy or deflect criticism. Such tactics were particularly evident in geopolitical contexts where aggressor nations sought to portray themselves as defenders of sovereignty or morality (Muñoz and Nuño, 2024). By systematically analyzing these patterns, AI tools provided insights into how language can be strategically deployed to manipulate public perception and align with broader disinformation campaigns.

3.17 Cross-referencing with semantic analysis

To ensure the robustness and reliability of findings, results derived from AI-driven analysis were cross-referenced with data obtained through semantic tools like the Attack-Index. This validation step was critical for enhancing the credibility of the conclusions drawn from the study. By aligning the thematic and emotional cues identified by AI models with the broader narrative patterns detected through semantic analysis, researchers achieved a cohesive understanding of how disinformation narratives are structured and disseminated. This cross-referencing methodology also minimized the risk of interpretive errors, ensuring that multiple analytical frameworks supported conclusions.

3.18 Unveiling biases

A significant component of the procedure involved the examination of biases embedded within disinformation narratives. AI algorithms were employed to uncover cognitive biases, such as confirmation bias or framing effects, which play a pivotal role in influencing target audiences' emotional and psychological responses. These biases were not only instrumental in shaping the narratives but also in enhancing their resonance and persuasiveness. For example, narratives framing international coalitions like NATO as aggressors capitalized on skepticism or distrust among specific demographic groups (Henschke, 2025). By unveiling these biases, the study provided a deeper understanding of the psychological underpinnings that make disinformation campaigns effective. Through these procedural steps, the study explored how manipulative narratives are constructed and perpetuated in cognitive warfare. Integrating AI tools with semantic methodologies provided a robust framework for identifying and analyzing the subtle yet impactful tactics used in modern disinformation efforts. This approach advanced the academic understanding of cognitive warfare and offered actionable insights for countering its psychological and geopolitical impacts.

3.19 Triangulation for robustness and reliability

Ensuring the validity and reliability of findings is paramount in any research endeavor, especially when analyzing complex phenomena such as manipulative narratives in cognitive warfare. This study employed triangulation by cross-verifying results derived from the Attack-Index and AI-driven analyses with expert reviews and independent datasets. This method not only bolstered the credibility of the findings but also mitigated potential biases stemming from the analytical tools themselves (Abu Arqoub, 2023; Sarwono, 2022). Experts in cognitive warfare and disinformation provided critical evaluations, helping to refine interpretations and ensure that conclusions aligned with real-world dynamics. Additionally, incorporating independent datasets further strengthened the study's foundation, allowing for comparative analysis and reducing the risk of over-reliance on any single data source.

4 Addressing ethical concerns

The ethical dimension of this research was a critical consideration, given the potential sensitivity and impact of findings related to disinformation and cognitive warfare. The study addressed several key ethical challenges:

4.1 Data privacy

Protecting individual privacy was a central focus, particularly when handling sensitive datasets from social media and other public platforms. To ensure compliance with ethical standards, anonymized data processing methods were employed. This approach safeguarded the identities of individuals while maintaining the integrity and relevance of the data for analysis (Huntsman et al., 2024).

4.2 Bias mitigation in AI tools

AI tools, while powerful, are inherently susceptible to biases that can skew analytical outcomes. To address this, findings from AI analyses were rigorously cross-referenced with expert reviews. This dual-layer validation minimized the influence of AI-induced biases and enhanced the reliability of the results (Nimmo and Flossman, 2024).

5 Results and analysis

This section presents the outcomes derived from utilizing the Attack-Index tool and AI-driven language models (LLMs). The focus is on identifying patterns, narratives, and emotional triggers inherent in cognitive warfare, particularly within the context of the Russia-Ukraine conflict.

The research on the Russian Federation's information space during two distinct periods—November 1, 2022 to March 5, 2023, and January to December 2024 utilized data from 4,000 Russian websites and 3,000 Telegram channels. This analysis revealed that nuclear threats remain a consistent and prominent theme within the Russian Federation's information landscape. It is deeply embedded in the primary propaganda narratives of the Russian Federation and reflected in the international segment of the Attack Index service's database.

Figure 1 highlights the dynamic nature of the nuclear threat mentioned in Russian propaganda. The graph captures the daily volume of publications on this topic, showing significant fluctuations and resonant peaks. Red lines mark instances where the Attack Index exceeded 30, signifying periods of heightened propaganda activity. The data indicates that October 2022 experienced the highest peak in information dynamics, with 807 daily mentions, representing the absolute maximum within the analyzed timeframe.

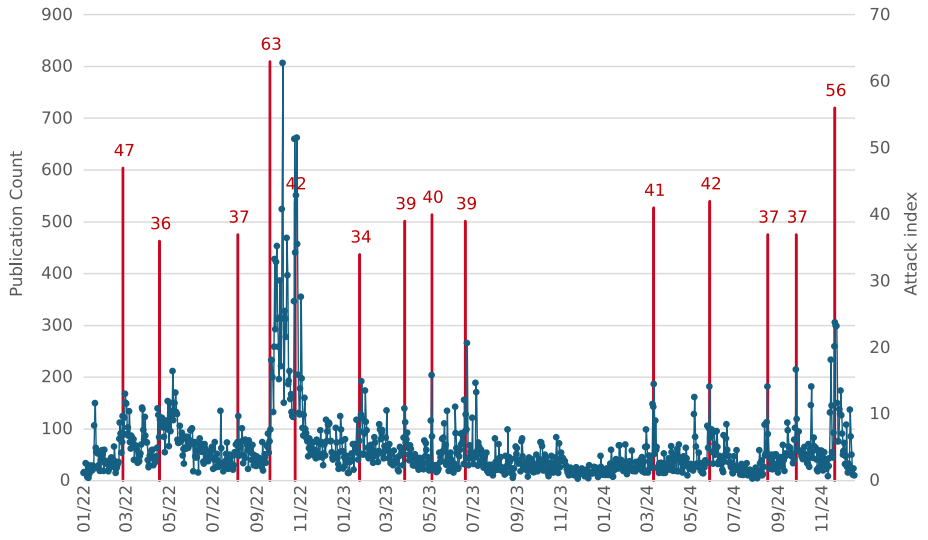

FIGURE 1
Information dynamics of mentions of possible nuclear strikes in Ukraine.

5.1 Attack-index findings

Using the clustering methodology outlined in the Appendix, five primary clusters were identified, each representing a significant thematic component of Russian disinformation narratives. These clusters highlight the strategic deployment of manipulative content in cognitive warfare.

5.1.1 "Special operation" and nuclear energy

This cluster encapsulated narratives associating the Ukraine conflict with potential nuclear disasters. Aimed predominantly at European audiences, these stories sought to exploit fears related to energy security and nuclear safety, amplifying anxiety about the conflict's broader ramifications.

5.1.2 "Kyiv authorities" and "dirty bomb"

Narratives in this cluster alleged that Ukraine was planning to deploy a "dirty bomb" to falsely implicate Russia in nuclear aggression. These claims aimed to delegitimize Ukraine while portraying Russia as a victim of unjust international accusations.

5.1.3 Russia's potential nuclear strike

This cluster focused on narratives emphasizing Russia's nuclear capabilities and potential readiness to deploy nuclear weapons. These stories were strategically crafted to deter Western military interventions and project an image of Russian dominance and resolve.

5.1.4 Nuclear programs and treaties

Discussions within this cluster revolved around international nuclear agreements, often highlighting alleged treaty violations or manipulations by Western powers. These narratives positioned Russia as an upholder of international norms, contrasting its actions with perceived breaches by adversaries.

5.1.5 Navy fleet and blackmail

This cluster centered on narratives involving the Russian naval fleet and its potential role in nuclear coercion or geopolitical blackmail. These stories emphasized Russia's strategic leverage and its ability to project power through its naval assets.

Figure 2 illustrates the emotional tone of publications about nuclear threat narratives in the Russian information space during the analyzed period. This emotional profiling, derived using machine learning techniques, highlights the asymmetry in the emotional dynamics of the narratives, characterized by a predominance of negative (harmful) content.

- Negative publications: out of the 68,206 posts analyzed, 35,974 posts (52.7%) were classified as negative. This significant proportion underscores the deliberate use of fear, distrust, and anxiety to manipulate public sentiment.
- Positive publications: positive narratives accounted for only 3,506 posts (5.1%) of the content, reflecting minimal emphasis on optimistic or reassuring messaging within these narratives.

Figure 3 shows the sources of analyzed posts by country where the country was determinable. TOT refers to Temporarily Occupied Territories of Ukraine. Unknown is used for analyzed posts where the country could not be accurately determined. Rest of World is a grouping of the remaining countries with post counts less than 350.

- Key narratives and themes: the attack-index tool revealed recurring motifs in Russian disinformation campaigns, such as nuclear threats, anti-NATO rhetoric, and victimization

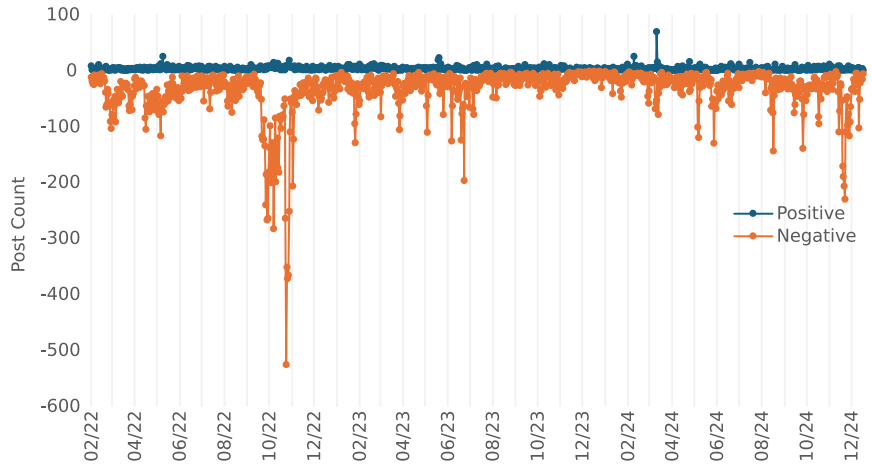

FIGURE 2
Sentiment analysis of the information dynamics of mentions of nuclear strikes.

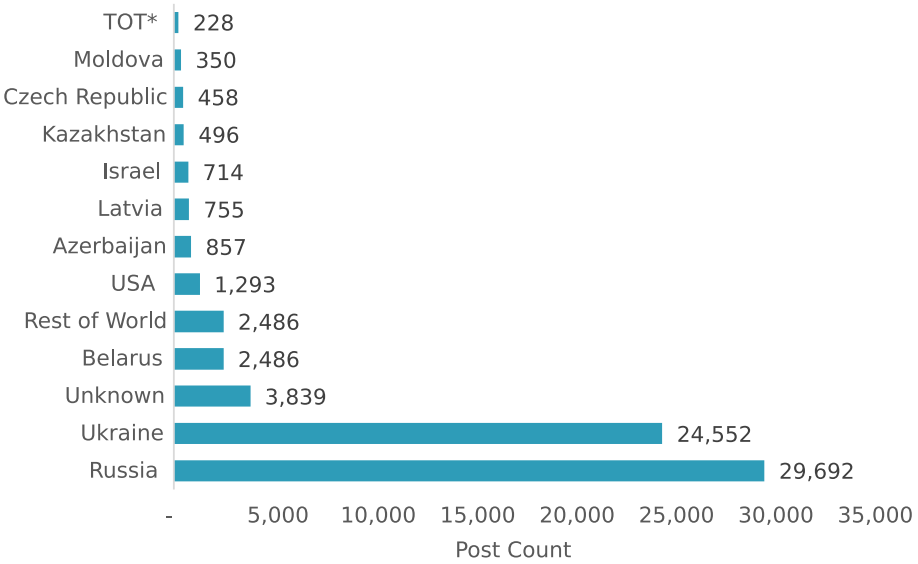

FIGURE 3
Post count by attributable country source.

narratives. These narratives were often synchronized with geopolitical events to maximize impact. For instance:

- Nuclear threat narratives: Russian state-controlled media repeatedly emphasized the potential for nuclear conflict to deter Western support for Ukraine (NATO Allied Command Transformation, 2024). These narratives spiked during NATO summits and military aid announcements.
- Anti-NATO rhetoric: disinformation campaigns framed NATO as an aggressor, accusing the alliance of threatening Russian sovereignty (Brusylovska and Maksymenko, 2023; Benkő and Biczók, 2024).

Temporal analysis revealed that shifts in narrative tone coincided with key geopolitical events. For example, during high-profile NATO meetings, Russian narratives shifted from reassurance to hostility, correlating with announcements of increased military aid to Ukraine (Monaghan, 2020). Supplementing the above analysis, data extracted from the Attack Index database demonstrates a strong correlation between nuclear threat narratives and significant geopolitical meetings. In particular, the Attack Index values leading up to and following high-profile events such as the Ramstein meetings, NATO and EU summits, and G20 sessions suggest a shift in Russian disinformation tactics. The data reveals that the topic of nuclear strikes often gains resonance before these events, with the Attack Index values peaking just ahead of these critical moments. This pattern reflects an intentional effort to preemptively shape global discourse around nuclear escalation and influence international decision-making regarding

military aid to Ukraine. Disinformation intensified during economic sanctions, leveraging themes of Western hypocrisy and Russian victimhood to bolster domestic support (NATO Allied Command Transformation, 2023; Deppe and Schaal, 2024). Russian state-controlled media persistently emphasizes nuclear threat narratives as a core component of its disinformation strategy. Analysis reveals that these narratives dominate both domestic and international segments of the Attack-Index database, reinforcing their importance in Russia's cognitive warfare arsenal between 2022 and 2024. This tactic underscores the Kremlin's reliance on nuclear rhetoric to manipulate perceptions and frame geopolitical discourse. The volume of nuclear threat mentions within Russian propaganda exhibits a resonant pattern, with significant peaks corresponding to pivotal geopolitical events. For example, during October 2022, mentions reached an absolute maximum of 807 instances per day, coinciding with heightened tensions in NATO-Ukraine relations. This temporal alignment indicates an intentional effort to exploit critical moments for maximum psychological impact.

Figure 4 illustrates the resonance values of the Attack Index on a temporal axis, plotted alongside information dynamics related to the Ramstein meetings, NATO and EU summits, and G20 sessions. The figure illustrates the strategic synchronization of disinformation campaigns with critical international milestones to maximize psychological and political impact. The data reveal a marked evolution in narrative synchronization. Note the blue and black columns that represent peaks in resonance values that occur predominantly before key international events, indicating a deliberate effort to shape narratives and public sentiment preemptively. This transition from reactive to proactive disinformation tactics underscores the adaptability of cognitive warfare strategies, particularly in leveraging nuclear rhetoric as a tool for psychological and strategic manipulation.

5.2 Correlation analysis and forecasting

The system conducted a correlation analysis to assess the relationship between mentions of nuclear threats and references to U.S. and NATO involvement in the Ukraine conflict. The analysis revealed that the highest correlation between these topics occurred in February 2021, with a Pearson correlation coefficient of 0.59. Forecast models projected increases in nuclear-related narratives, anticipating 600–700 daily mentions by March 2023. Following an escalation in nuclear threat rhetoric in October 2022, the correlation began to rise again, stabilizing at 0.45 by March 2023. This indicates a sustained link between discussions of nuclear threats and mentions of U.S. and NATO actions in the conflict. To complement this analysis, a forecasting model was developed using the Attack-Index system. The model projected a decline in nuclear threat mentions in early March 2023, followed by an increase, reaching 600–700 daily publications by the end of the month. This prediction utilized historical trends in mentions of nuclear threats alongside references to U.S. and NATO involvement.

5.3 AI-driven analysis

Purpose and Scope AI tools, including LLMs, were deployed to uncover nuanced manipulative tactics, such as sarcasm, euphemisms, and framing techniques. This approach enhanced the understanding of emotional and psychological appeals embedded in disinformation narratives. Cross-verification results from the Attack-Index tool were corroborated with AI-driven analyses to validate key insights. This integration provided a robust framework for understanding narrative construction and dissemination in cognitive warfare (Deppe and Schaal, 2024).

5.4 Key findings

Sentiment analysis of the nuclear threat narratives in Figure 2 highlights their predominantly negative tone. Approximately 52.7% of mentions expressed negative sentiment, compared to only 5.1% classified as positive. This deliberate focus on fear-based messaging exemplifies Russia's strategy of leveraging existential threats to undermine public confidence and provoke anxiety within target audiences. Linguistic Patterns: both tools highlighted

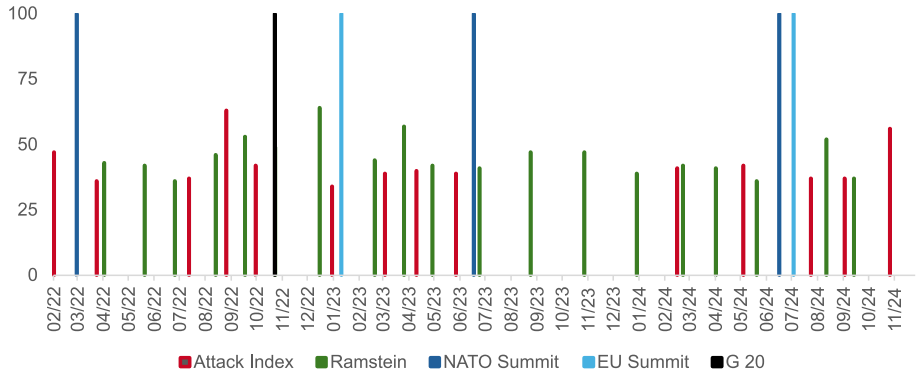

FIGURE 4
Temporal resonance values of the attack index over time, highlighting peaks in narrative activity preceding key geopolitical events.

the consistent use of emotionally charged language and logical fallacies, such as false equivalencies, to influence public sentiment (Huang et al., 2024; Grindrod, 2024). Psychological Exploitation: narratives exploited cognitive biases, including confirmation bias and availability heuristics, ensuring resonance with target audiences (Rabb et al., 2021; Monaghan, 2020). Narratives targeting fear and moral outrage effectively shape public opinion, leveraging psychological triggers to enhance resonance and spread (Vosoughi et al., 2018). Fear and Anger: Russian narratives leveraged existential threats and moral outrage to evoke fear and anger. For instance, nuclear threat narratives targeted Western audiences to discourage military interventions (Weissmann et al., 2021). Moral Outrage: emotional framing, such as claims of NATO's moral hypocrisy, was amplified before pivotal geopolitical moments, influencing public perception (Grindrod, 2024). Euphemisms and Sarcasm: AI tools detected subtle manipulative language that obscured aggression while projecting victimhood. For example, references to "peacekeeping" operations masked overt military actions (Muñoz and Nuño, 2024). Framing Techniques: the strategic use of framing presented Russia's actions as defensive while portraying NATO as a destabilizing force (Henschke, 2025). Russia reframes its aggression by portraying itself as a defender and victim, shifting blame to Western support for Ukraine. This narrative strategy inverts the roles of aggressor and victim to justify its actions and weaken global opposition (Turska, 2024). Longitudinal Trends: analysis revealed that narrative themes evolved, adapting to audience receptivity and geopolitical dynamics (Henschke, 2025; Van der Klaauw, 2023).

5.5 Geopolitical synchronization

Russian disinformation narratives display a high degree of synchronization with major geopolitical events, strategically amplifying their psychological impact. AI-driven analyses identified significant increases in victimhood narratives following Western sanctions, with Russia framed as unfairly targeted by the international community (Plaza et al., 2023). Temporal analysis using AI tools revealed that shifts in disinformation narratives were closely tied to real-world events. Claims of Western duplicity were prominently emphasized before key international conferences, leveraging public distrust to undermine alliances. Analysis revealed narrative spikes coinciding with NATO summits and major sanctions announcements, as detailed in the timeline of narrative resonance (Figure 4). These alignments illustrate the calculated synchronization of disinformation efforts with critical geopolitical events. By 2024, this synchronization had become increasingly proactive, with disinformation campaigns peaking in advance of these events to shape public opinion preemptively. This strategic alignment allowed adversaries to control the narrative and influence decision-making processes ahead of time. The study further demonstrated that disinformation narratives are deliberately synchronized with specific geopolitical developments to enhance their resonance and psychological effect. For instance, during NATO summits, anti-NATO rhetoric spiked, portraying Western actions as deliberate provocations against Russian sovereignty (Benkő and Biczók, 2024). These findings underscore the sophisticated tactics employed in cognitive warfare, highlighting the need for enhanced detection and countermeasures to mitigate their impact.

6 Discussion

6.1 Key contributions to literature

The findings of this study contribute significantly to the evolving body of research on cognitive warfare and disinformation, offering novel insights into the mechanisms and impacts of manipulative narratives in modern geopolitical conflicts. By identifying key themes such as nuclear threats, anti-NATO rhetoric, and victimization, the study highlights their centrality in Russian disinformation campaigns. These narratives were not randomly deployed but strategically aligned with significant geopolitical events to amplify their psychological impact and influence public perception (Monaghan, 2020).

6.2 Psychological triggers and cognitive biases

Emotional triggers—such as fear, anger, and moral outrage—were systematically leveraged to manipulate public sentiment and polarize target audiences. These emotions were not merely incidental; they were deliberately crafted to exploit cognitive biases like confirmation bias and availability heuristics, which inherently predispose individuals to accept information that aligns with pre-existing beliefs or recent memories (Plaza et al., 2023; Grindrod, 2024). By examining the strategic use of these biases, the study advances theoretical understandings of how psychological vulnerabilities can be weaponized in cognitive warfare. The findings expand the theoretical understanding of cognitive warfare by highlighting its psychological and technological dimensions, particularly the interplay between emotional triggers and narrative framing (Van der Klaauw, 2023).

6.3 Advanced manipulative strategies

The application of AI-driven tools illuminated nuanced manipulative tactics employed in disinformation campaigns. These included the use of euphemisms to obscure intent, sarcasm to undermine credibility, and framing techniques that redefined aggression as victimhood. These advanced rhetorical strategies underscore the sophistication of modern cognitive warfare, where language becomes a tool for psychological subversion (Huang et al., 2024; Muñoz and Nuño, 2024).

6.4 Temporal and strategic dynamics

The findings revealed a clear synchronization of narrative shifts with critical geopolitical developments, such as NATO summits and announcements of military aid. For example, during NATO meetings, narratives shifted from passive to hostile tones,

leveraging the event's visibility to maximize strategic impact. This temporal alignment underscores the calculated nature of cognitive operations, aiming to influence public opinion and policymaking during key geopolitical milestones (Benkő and Biczók, 2024).

6.5 Methodological innovations

The methodological framework of this study employs a dual-layered approach, integrating semantic analysis with AI-driven techniques, exemplified by the Attack Index. This innovative combination enables a nuanced understanding of narrative construction and thematic evolution while providing real-time insights into the emotional resonance of manipulative narratives. By identifying psychological triggers, such as fear and moral outrage embedded within disinformation campaigns, the research advances theoretical perspectives and offers practical roadmaps for crafting targeted counter-narratives and preemptive strategies. Through the analysis of historical data and current narrative dynamics, the study demonstrates the predictive capabilities of advanced analytics in identifying and mitigating disinformation. This proactive approach underscores the critical importance of real-time monitoring systems and international collaboration in addressing the transnational challenges posed by cognitive warfare, ensuring resilience in an evolving threat landscape. The integration of semantic analysis tools like the Attack Index with AI-driven methodologies provides a robust framework for analyzing disinformation narratives. Semantic tools systematically identified recurring themes and emotional triggers, while AI models unveiled deeper layers of manipulation, such as logical fallacies and rhetorical devices. Cross-verification through triangulation ensured the robustness and reliability of findings, mitigating potential biases inherent in automated tools and enhancing the validity of results (Abu Arqoub, 2023; Sarwono, 2022). By bridging qualitative and quantitative methodologies, the study addresses a critical gap in cognitive warfare research. It demonstrates the effectiveness of combining human-centered analysis with AI technologies to capture both the psychological and technological dimensions of manipulative narratives. This integration advances academic understanding and provides actionable insights for countering disinformation and fortifying defenses against cognitive warfare (Deppe and Schaal, 2024; Henschke, 2025).

6.6 Further research

Further research is needed to analyze disinformation narratives' emotional and psychological impact, focusing on the role of specific emotional triggers like hope, despair, or indignation in shaping public perception (Hoyle et al., 2023; Rabb et al., 2021). Examining the evolution of cognitive warfare strategies across different conflicts and cultural contexts can provide insights into the adaptability of manipulative narratives and their long-term effects on societal trust (Brusylovska and Maksymenko, 2023). Combining perspectives from psychology, linguistics, political science, and computer science can enrich the understanding of cognitive warfare, fostering innovative methodologies for analyzing and countering its effects (Grindrod, 2024; Henschke, 2025). Collaboration between academia, governments, and international organizations is essential for developing standardized frameworks to address the complex challenges posed by cognitive warfare. Future research should focus on crafting actionable policies that enhance societal resilience and foster international cooperation (Marsili, 2023; NATO Allied Command Transformation, 2023).

7 Practical implications

The findings of this study underscore the sophistication and strategic nature of modern cognitive warfare, offering significant practical implications for policymakers, organizations, and governments seeking to counter disinformation effectively. Findings from network visualizations (Appendix Figure 5) and sentiment trends (Figure 2), provide a roadmap for developing real-time monitoring systems and counter-narrative strategies tailored to evolving disinformation trends. By integrating semantic and AI-driven analyses, the research provides actionable insights into the detection, prevention, and mitigation of manipulative narratives. The deployment of tools like the Attack-Index in real-time monitoring systems emerges as a cornerstone for addressing the complex challenges posed by cognitive warfare in contemporary geopolitical contexts (Huang et al., 2024; Weissmann et al., 2021).

7.1 Detection and prevention

Enhanced analytical tools, such as the Attack-Index and AI-based algorithms, play a crucial role in identifying emotional triggers, thematic clusters, and shifts in disinformation narratives. These tools allow policymakers and practitioners to detect manipulative strategies early, preempting their psychological and societal impacts. For example, by identifying spikes in nuclear threat narratives or anti-NATO rhetoric, these tools can provide early warnings of targeted disinformation campaigns designed to exploit geopolitical tensions (Huang et al., 2024). Real-time detection of emotional triggers, such as fear or moral outrage, also enables governments and organizations to tailor timely responses, mitigating the resonance and spread of harmful narratives (Plaza et al., 2023).

7.2 Strategic response

The study highlights how narrative synchronization with critical geopolitical events—such as NATO summits, military aid announcements, or sanctions—enhances the psychological impact of disinformation. By analyzing the alignment between these events and narrative shifts, policymakers can develop targeted countermeasures. Strategic responses, such as public awareness campaigns and transparent communication strategies, can disrupt the momentum of manipulative narratives before they reach their peak impact. For instance, preemptive public disclosures and fact-checking initiatives during key geopolitical milestones can diminish the credibility of disinformation and foster public resilience against cognitive manipulation (Nimmo and Flossman, 2024).

7.3 Real-time monitoring and decision support

The integration of tools like the Attack-Index into real-time monitoring systems provides an invaluable resource for decision-makers. These systems can deliver actionable intelligence by tracking disinformation patterns and emotional appeals across digital platforms, enabling a proactive approach to cognitive warfare. For example, governments and international organizations could deploy real-time monitoring dashboards to analyze the spread of narratives during high-stakes negotiations or elections, informing strategic decisions and communication policies (NATO Allied Command Transformation, 2024; Index Systems Ltd., 2023).

7.4 International collaboration

The transnational nature of cognitive warfare necessitates a coordinated response that transcends national borders. This study underscores the need for international collaboration in developing standardized frameworks and tools for countering disinformation. Collaborative efforts among governments, academic institutions, and technology companies can enhance the effectiveness of countermeasures, such as joint fact-checking initiatives, shared monitoring platforms, and collective policy responses. For example, NATO's efforts to integrate cognitive warfare countermeasures into its broader strategic framework demonstrate the potential for multinational cooperation in addressing this evolving threat (NATO Allied Command Transformation, 2023; Nimmo and Flossman, 2024).

7.5 Leveraging advanced technologies

Future research and practical applications should explore the potential of advanced technologies, such as Generative Adversarial Networks (GANs) and deepfake detection tools, in both propagating and mitigating manipulative narratives. The Appendix documents the application of clustering algorithms that lays a foundation for future research on scalable AI-driven methodologies to combat cognitive warfare and disinformation. These technologies can either enhance or counteract cognitive warfare, depending on their governance and application. Policymakers must prioritize the development of ethical, efficient and effective AI workflow tools such as Opus that can identify sophisticated disinformation tactics, such as deepfakes or synthetic media, ensuring they do not exacerbate the challenges of cognitive warfare (Fagnoni et al., 2024).

7.6 Building societal resilience

Beyond technological interventions, fostering societal resilience against cognitive warfare is critical. Media literacy programs and public awareness campaigns can empower individuals to recognize and resist manipulative narratives. These initiatives should target vulnerable populations, such as youth and digitally marginalized communities, who are often disproportionately affected by disinformation. By promoting critical thinking skills and digital literacy, governments and organizations can reduce the societal impact of cognitive warfare, enhancing overall resilience (Sarwono, 2022; Wallenius, 2023).

8 Limitations

Despite this study's comprehensive framework and robust methodologies, several limitations must be acknowledged to contextualize the findings and inform future research. While extensive, the datasets used in this study are not immune to biases inherent in the sources from which they were derived. For instance, Russian state-controlled and social media platforms may reflect selective or exaggerated narratives intended to manipulate perceptions, potentially skewing the analysis. Additionally, the reliance on publicly available data excludes covert or less-detectable disinformation campaigns, which could provide further insights into manipulative strategies. These limitations highlight the need for more diverse and representative datasets to comprehensively understand cognitive warfare. Although advanced AI tools like Large Language Models (LLMs) and the Attack-Index offer significant advantages in detecting and analyzing disinformation, they have flaws. AI systems are susceptible to biases introduced during training, which may influence the interpretation of narratives and emotional triggers. Moreover, these tools may struggle to capture complex cultural nuances or context-specific manipulations, limiting their applicability across diverse geopolitical environments. The inability of current AI models to fully account for sarcasm, irony, or deeply embedded cultural references may result in partial or incomplete analyses. Measuring the long-term effects of cognitive warfare on societal trust, political stability, or international relations presents a significant challenge. The psychological and societal impacts of disinformation often manifest over extended periods, making it difficult to establish causal links between specific narratives and broader outcomes. Additionally, the interplay of multiple factors—economic conditions, political developments, and media ecosystems—further complicates the measurement of disinformation's long-term effects. The rapidly evolving nature of cognitive warfare presents another limitation. While this study captures current strategies and tools, adversaries continuously develop more sophisticated methods, such as Generative Adversarial Networks (GANs) and more advanced AI-driven disinformation techniques. As a result, the findings may have limited applicability as the tactics of cognitive warfare evolve. Future studies must remain adaptable, employing iterative methodologies to account for emerging technologies and shifting geopolitical landscapes. Ethical considerations, such as ensuring data privacy and avoiding the misuse of findings, imposed certain restrictions on the scope of this research. While these measures were necessary to uphold ethical standards, they may have limited access to sensitive or proprietary data that could have enriched the analysis. Operationally, integrating multiple analytical tools required significant computational and logistical resources, which could constrain scalability in broader applications.

9 Conclusion

This study provides a detailed exploration of cognitive warfare, focusing on its mechanisms, impacts, and strategic dimensions within the context of the Russia-Ukraine conflict. Employing a robust methodological framework that integrates semantic analysis tools like the Attack-Index and AI-driven approaches, the research sheds light on constructing and disseminating manipulative narratives. The findings illustrate the sophistication of modern cognitive warfare, characterized by emotional exploitation, narrative synchronization, and the strategic use of technological tools. This research contributes to the broader understanding of cognitive warfare as a critical element of contemporary hybrid conflicts by offering actionable insights into countering disinformation. In summary, this research makes significant contributions to the understanding of cognitive warfare, particularly its application during the Russia-Ukraine conflict.

The paper's key contributions include:

1. Establishing nuclear rhetoric as a strategic tool within manipulative narratives, highlighting its synchronization with geopolitical events such as NATO summits and military aid meetings.
2. Introducing an innovative analytical framework that combines semantic analysis and AI-driven methodologies, enhancing the detection and prediction of disinformation narratives.
3. Providing actionable insights for crafting countermeasures, with a focus on the importance of real-time monitoring systems and international collaboration to address the transnational challenges posed by cognitive warfare.

By addressing these dimensions, the research bridges critical gaps in understanding cognitive warfare's mechanisms, impacts, and countermeasures, offering both theoretical and practical advancements for mitigating disinformation in contemporary conflicts.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary material.

Author contributions

AP: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. DL: Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. ES-T: Data curation, Investigation, Writing – original draft, Writing – review & editing. PK: Formal analysis, Software, Visualization, 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.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. To analyze manipulative narratives within cognitive warfare using AI-driven tools such as semantic analysis and pre-trained language models. Specifically, these tools were employed to identify subtleties like sarcasm, euphemisms, and framing techniques, as well as to enhance the detection of cognitive biases embedded within disinformation campaigns. Additionally, AI-supported methodologies were utilized to support the linguistic analysis, temporal correlation, and thematic clustering of data, contributing to the robustness and depth of the research findings.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1566022/full#supplementary-material

References

Abu Arqoub, O. (2023). Reputation themes from communication perspective: a qualitative systematic review. *J. Assoc. Arab Univ. Res. High. Educ.* 43, 329–351. doi: 10.36024/1248-043-003-018

Benkő, G., and Biczók, G. (2024). The Cyber Alliance Game: how alliances influence cyber-warfare. *arXiv preprint* arXiv:2410.05953. doi: 10.48550/arXiv.2410.05953

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press: New York, NY, USA. doi: 10.1093/oso/9780190923624.001.0001

Bilal, A. (2021). H*ybrid Warfare - New Threats, Complexity, and 'Trust' as the Antidote*. NATO Review. Available online at: https://www.nato.int/docu/review/articles/2021/11/30/hybrid-warfare-new-threats-complexity-and-trust-as-the-antidote/index.html (Accessed December 17, 2024).

Bradshaw, S., and Howard, P. N. (2019). "The global disinformation order: 2019 global inventory of organised social media manipulation" in *Computational Propaganda Project, Working Paper 2019.2* (Oxford: Oxford Internet Institute; University of Oxford). Available online at: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf

Bruns, A., and Snee, H. (2022). *How to Visually Analyse Networks Using Gephi*. SAGE Publications, Limited: London, UK. doi: 10.4135/9781529609752

Brusylovska, O., and Maksymenko, I. (2023). Analysis of the media discourse on the 2022 war in ukraine: the case of russia. *Reg. Sci. Policy Pract.* 15, 222–235. doi: 10.1111/rsp3.12579

Buzan, B., and Waever, O. (2003). *Regions and Powers: The Structure of International Security, Volume 226*. Cambridge University Press: Cambridge, UK.

Deppe, C., and Schaal, G. S. (2024). Cognitive warfare: a conceptual analysis of nato act. *Front. Big Data* 7:1452129. doi: 10.3389/fdata.2024.1452129

Digital Forensic Research Lab (2023). *Russian War Report: DFRLab Releases Investigations on Russian Info Ops Before and After the Invasion*. Atlantic Council. Available online at: https://dfrlab.org/2023/02/25/russian-war-report-dfrlab-releases-investigations-on-russian-info-ops-before-and-after-the-invasion/

DiResta, R. (2020). The supply of disinformation will soon be infinite. *The Atlantic*. Available online at: https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400 (Accessed October 29, 2024).

Fagnoni, T., Mesbah, B., Altin, M., and Kingston, P. (2024). Opus: a large work model for complex workflow generation. *arXiv preprint arXiv:2412.00573*. doi: 10.48550/arXiv.2412.00573

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., et al. (2024). Biased AI can influence political decision-making. *arXiv preprint arXiv:2410.06415*. doi: 10.48550/arXiv.2410.06415

Gambhir, H. (2016). *The Virtual Caliphate: Isis's Information Warfare*. Research report, Institute for the Study of War: Washington, DC, USA.

Grindrod, J. (2024). Linguistic intentality in large language models: applications for countering disinformation. *Synthese* 204:71. doi: 10.1007/s11229-024-04723-8

Henschke, A. (2025). *Cognitive Warfare: Grey Matters in Contemporary Political Conflict*. Routledge: London, UK. doi: 10.4324/9781003126959

Hoyle, A., Wagnsson, C., van den Berg, H., Doosje, B., and Kitzen, M. (2023). Cognitive and emotional responses to russian state-sponsored media narratives in international audiences. *J. Media Psychol.* 35, 362–374. doi: 10.1027/1864-1105/a000371

Huang, A., Pi, Y. N., and Mougan, C. (2024). "Moral persuasion in large language models: evaluating susceptibility and ethical alignment," in *AdvML-Frontiers Workshop* (Vancouver, BC: NeurIPS 2024). doi: 10.48550/arXiv.2411.11731

Huntsman, S., Robinson, M., and Huntsman, L. (2024). *Prospects for Inconsistency Detection Using Large Language Models and Sheaves*. Research Report, OpenAI Research.

Index Systems Ltd (2022). *Analysis of Narratives in the Russian Information Space: Nuclear Threats*. AttackIndex Report (Kyiv).

Index Systems Ltd (2023). *UKR039 Cluster Report: Correlational Analysis of Russian Information Space*. AttackIndex Report (Kyiv).

Jensen, B., and Ramjee, D. (2023). *Beyond Bullets and Bombs: The Rising Tide of Information War in International Affairs*. Washington, DC: Report, Center for Strategic and International Studies. Available online at: https://www.csis.org/analysis/beyond-bullets-and-bombs-rising-tide-information-war-international-affairs

Lande, D., and Shnurko-Tabakova, E. (2019). Osint as a part of cyber defense system. *Theor. Appl. Cybersecur.* 1, 103–108. doi: 10.20535/tacs.2664-29132019.1.169091

Lande, D. V. (2024). *OSINT in Cybersecurity: Study Guide*. Engineering LLC, Kyiv, Ukraine.

Lewandowsky, S., Ecker, U. K., and Cook, J. (2017). Beyond misinformation: understanding and coping with the "post-truth" era. *J. Appl. Res. Mem. Cogn.* 6, 353–369. doi: 10.1016/j.jarmac.2017.07.008

Marsili, M. (2023). Guerre à la carte: Cyber, information, cognitive warfare and the metaverse. *ACIG* 2, 1–15. doi: 10.60097/ACIG/162861

Marwick, A., and Lewis, R. (2017). *Media Manipulation and Disinformation Online*. New York, NY: Data & Society Research Institute. Available online at: https://datasociety.net/library/media-manipulation-and-disinformation-online/

Monaghan, S. (2020). Countering hybrid warfare: So what for the future joint force? *PRISM* 8, 83–98. doi: 10.2307/resrep26552

Muñoz, F. J., and Nuño, J. C. (2024). Winning opinion: following your friends' advice or that of their friends? *arXiv preprint* arXiv:2411.16671. doi: 10.48550/arXiv.2411.16671

NATO Allied Command Transformation (2023). *Mitigating and Responding to Cognitive Warfare*. Norfolk, VA: Allied Command Transformation. Available online at: https://www.act.nato.int/articles/mitigating-and-responding-to-cognitive-warfare

NATO Allied Command Transformation (2024). *Allied Command Transformation Develops the Cognitive Warfare Concept to Combat Disinformation and Defend Against "Cognitive Warfare"*. Norfolk, VA: NATO ACT. Available online at: https://www.act.nato.int/article/cogwar-concept/?utm_source=chatgpt.com"https://www.act.nato.int/article/cogwar-concept/ (Accessed December 17, 2024).

Nguyen, T. N. (2023). Accelerated cognitive warfare via the dual use of large language models. *Preprints*. doi: 10.20944/preprints202312.2279.v1

Ngwainmbi, E. K. ed. (2024). *Social Media, Youth, and the Global South: Comparative Perspectives*. Palgrave Macmillan Cham, Cham, 1 edition. Literature, Cultural and Media Studies Series. doi: 10.1007/978-3-031-41869-3

Nimmo, B., and Flossman, M. (2024). *Influence and Cyber Operations: An Update*. San Francisco CA: OpenAI. Available online at: https://openai.com/safety/influence-and-cyber-operations-update-oct-2024 (Accessed October 29, 2024).

Orinx, K., and Struye de Swielande, T. (2022). "China and cognitive warfare: why is the West losing?," in *Cognitive Warfare: The Future of Cognitive Dominance*, eds. B. Claverie, B. Prébot, N. Buchler, and F. Du Cluzel (Neuilly-sur-Seine: NATO Science and Technology Organization, Collaboration Support Office), 1–6. doi: 10.48550/hal-03635930

Pennycook, G., Bear, A., Collins, E. T., and Rand, D. G. (2020). The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manage. Sci.* 66, 4944–4957. doi: 10.1287/mnsc.2019.3478

Plaza, F., Sotelo Monge, M., and Ordi, H. (2023). "Towards the definition of cognitive warfare and related countermeasures: a systematic review," in *Proceedings of the 2023 International Conference on Cybersecurity, Situational Awareness and Social Media (CyberSA 2023)*, –7. doi: 10.1145/3600160.3605080

Rabb, N., Cowen, L., de Ruiter, J. P., and Scheutz, M. (2021). Cognitive cascades: how to model (and potentially counter) the spread of fake news. *PLoS ONE* 16:e0246234. doi: 10.1371/journal.pone.0261811

Sarwono, J. (2022). *Quantitative*, qualitative, and mixed method research methodology.

Splidsboel Hansen F. (2021). When Russia wages war in the cognitive domain. *J. Slavic Milit. Stud.* 34, 181–201. doi: 10.1080/13518046.2021.1990562

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* 9:5233. doi: 10.1038/s41598-019-41695-z

Tucker, J. A., Theocharis, Y., Roberts, M. E., and Barberá, P. (2017). From liberation to turmoil: social media and democracy. *J. Democr.* 28, 46–59. doi: 10.1353/jod.2017.0064

Turska, K. (2024). The risks to the world from Russia's nuclear blackmail. *The Post*. Available online at: https://www.thepost.co.nz/nz-news/360507423/risks-world-russias-nuclear-blackmail (Accessed October 29, 2024).

van der Klaauw, C. (2023). Cognitive warfare: NATO's 21st-century game changer. *Three Swords*. 39, 97–99. Availablle online at: https://www.jwc.nato.int/newsroom/three-swords/#

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

Wallenius, C. (2023). *Do Hostile Information Operations Really Have the Intended Effects? A Literature Review*. Karlstad: Swedish Defence University; Department of Security, Strategy, and Leadership. Available online at: https://www.diva-portal.org/smash/get/diva2:1661333/FULLTEXT01.pdf

Weissmann, M., Nilsson, N., Palmertz, B., and Thunholm, P. (2021). *Hybrid Warfare: Security and Asymmetric Conflict in International Relations*. I.B. Tauris: London, UK. doi: 10.5040/9781788317795

Wu, F. Y. (1982). The potts model. *Rev. Mod. Phys.* 54, 235–268. doi: 10.1103/RevModPhys.54.235

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 24 September 2025
DOI 10.3389/frai.2025.1616447

OPEN ACCESS

EDITED BY
Ludmilla Huntsman,
Cognitive Security Alliance, United States

REVIEWED BY
J. D. Maddox,
George Mason University, United States

*CORRESPONDENCE
Paul Thompson
8paul.thompson@gmail.com

RECEIVED 22 April 2025
ACCEPTED 29 August 2025
PUBLISHED 24 September 2025

CITATION
Thompson P and Guillory S (2025) The history of the semantic hacking project and the lessons it teaches for modern cognitive security.
Front. Artif. Intell. 8:1616447.
doi: 10.3389/frai.2025.1616447

The history of the semantic hacking project and the lessons it teaches for modern cognitive security

Paul Thompson[1]* and Sean Guillory[2]

[1]Dartmouth College, Hanover, NH, United States, [2]MAD Warfare, Portland, OR, United States

The Semantic Hacking Project ran from 2001 to 2003. It focused on how information systems (and the human decisions shaped by them) could be exploited through attacks not on code or infrastructure, but on meaning. This work is relevant to contemporary cognitive security concerns in the face of today's information space. The work provides insight into the key question of how people come to hold the beliefs which they do. The project anticipated many of today's challenges (disinformation campaigns, social media manipulation, AI-generated narratives) not just in technical terms, but in philosophical and linguistic terms. At the heart of its concern was a simple but powerful question: What happens when you can manipulate the inputs to a person's belief system without the person knowing it? This question has only grown more urgent in an era of generative AI, large language models (LLMs), and algorithmically amplified influence.

KEYWORDS

artificial intelligence, generative AI, large language models, countermeasures, cognitive security

1 Introduction: definitions as weapons

While this special issue of *Frontiers Science* will present a variety of definitions for "cognitive security," we want to offer a conceptual framing that explains why this domain matters so deeply for national security and warfare. For us, cognitive security is not just a new subset of cybersecurity or psychological operations; it is a foundational way to understand modern conflict itself.

Warfare has never been purely about kinetic force. While what happens in the physical domain is possibly the most extreme physical and psychological experience known to man, most of the real contest happens in the cognitive domain: the space of interpretation, belief, perception, and judgment. It is in this contested arena of ideas, narratives, and sense-making (the outcome of which is decided away from psychical battlefield concerns by people not physically present in the battlefield themselves) that who really won or lost a conflict is resolved. Whether a missile hits its target, whether a nation complies with sanctions, or whether an election's outcome is accepted by its population, all of these hinge less on what physically happens, but more on what people believe has happened and why (Perception is/as Reality).

War is itself is a communicative act where every kinetic operation sends a message, whether that message is intentional or not. A successful airstrike or special operation can achieve all of its tactical objectives and still fail strategically if the resulting narrative (from the adversaries on the battlefield, the adversaries off the battlefield, and all the civilian audiences in between) does not "hit" as planned. The moment action is taken, it becomes part of a communicative sequence, interpreted by different audiences through varying lenses. In this

sense, modern war is not just about what is done, but about what it means to adversaries, allies, civilian populations, and the broader international system.

It's understood if the reader is having difficulty understanding the what, the where, and how the cognitive domain works. It is like trying to understand the physics and biology of a Looney Tunes universe that works off of the rules of imagination itself. But trying to measure it and understand it is a must to anyone who is serious about figuring out how to win wars. People often treat fiction as fact if it aligns with their worldview; realities can be ignored; and symbols can override material truths.

And we would argue that in these fights that happen in the cognitive domain, the battles over definitions themselves may be the most important because whatever is the definition of gender, genocide, A.I., and the like that are agreed upon by the participants and audience hold the power in defining a situation and what people think should happen next. To correct the most famous quote from Dune, "He who controls the definitions, controls the universe."

Now with that contextual layout, we get to the importance of the central concept of our paper: semantic hacking. Beginning in 2001, a project at Dartmouth College's Institute for Security Technology Studies launched one of the earliest systematic efforts to explore how information systems could be attacked not by disrupting their hardware or code, but by manipulating the meaning of the data they carry. This effort, known as the Semantic Hacking Project, sought to understand and model the kinds of attacks that target human judgment, language, and interpretation. The project introduced the concept of "cognitive attacks" and explored a range of countermeasures aimed at defending decision-making processes from deliberate semantic distortion.

What makes this work relevant today is not just its prescience to discussions about modern cognitive security concerns in the face of today's information space, but its insight into the key question of how people come to believe what they believe. The project anticipated many of today's challenges (disinformation campaigns, social media manipulation, AI-generated narratives) not just in technical terms, but in philosophical and linguistic terms. At the heart of its concern was a simple but powerful question: What happens when you can manipulate the inputs to a person's belief system without them knowing it? This question has only grown more urgent in an era of generative AI, large language models (LLMs), and algorithmically amplified influence. For example, consider current debates over whether an artificial intelligence system is "sentient." The core of that discussion often comes down not to the system's actual behavior, but to the *definition* of "sentient" that is adopted. Change the definition, and the outcome of the debate changes accordingly. The same applies to politically or morally loaded concepts like "life," "gender," "sovereignty," or "genocide." In each case, the battle over perception begins with a battle over language and categories. Once a term is defined in a particular way, the logical conclusions that follow often feel self-evident to those operating within that frame. But as the Semantic Hacking Project emphasized, those frames are often up for grabs, and can be subtly or overtly manipulated with the help of AI across multiple media channels at scale. Again, these were insights from almost 25 years ago.

This paper will explore the history of the Semantic Hacking Project and the intellectual trajectory it initiated, including the workshops and academic developments that followed. We will then contextualize those early insights in light of modern cognitive security challenges, particularly those raised by artificial intelligence. In doing so, we hope to demonstrate that many of the threats being discussed today are not entirely new. Rather, they are the modern instantiations of conceptual battles that have been quietly brewing for decades. We share all this because we believe that understanding this somewhat forgotten line of research could offer a useful foundation for the urgent work ahead.

2 The semantic hacking project: origins and intellectual contributions

The Semantic Hacking Project was one of the early efforts of Dartmouth College's Institute for Security Technology Studies (ISTS), now known as the Institute for Security, Technology, and Society. Running from 2001 to 2003, the project was ahead of its time in its focus on how information systems (and the human decisions shaped by them) could be exploited through attacks not on code or infrastructure, but on meaning.

The term "semantic attack" had previously been introduced by Libicki (1994), who categorized computer network attacks into three types: physical (targeting hardware), syntactic (targeting system functionality, e.g., with viruses or worms), and semantic. Semantic attacks aimed not to break systems, but to manipulate decision-making whether by a human, a software agent, or an organization.

Building on this conceptual foundation, the Dartmouth team (led by George Cybenko and including Annarita Giani and co-author of this paper Paul Thompson) launched a project that would reframe these semantic threats as "cognitive attacks," emphasizing the human interpretive processes at risk. While primarily conceptual, the project did culminate in a working prototype of one of the proposed countermeasures. More significantly, it produced one of the first structured taxonomies of cognitive threats and countermeasures in the information domain. Initial results were presented at workshops beginning in 2002, and the project's first journal publication followed shortly thereafter (Cybenko et al., 2002a; Cybenko et al., 2002b; Thompson, 2003). A more comprehensive treatment was published in a 2004 volume of *Advances in Computers* (Cybenko et al., 2004). The team's work remains one of the earliest and most comprehensive treatments of semantic and cognitive hacking in national security contexts.

2.1 Reframing the problem: from syntactic disruption to cognitive exploitation

Unlike traditional cyberattacks at the time focusing on breaking systems, cognitive attacks work by distorting how humans perceive reality and make decisions. Rather than taking down a server or corrupting data, a cognitive attack shifts the interpretation of information. The attacker does not need to control the system, they only need to control the conclusions the system's users draw from it. In the realm of cyber, avid practitioners of "social engineering" had applied experience that led them to such insights but in terms of structured academic studies on the subject, it was The Semantic Hacking Project that categorized potential countermeasures into two types based on the structure of the

information environment: those designed for single-source contexts and those designed for environments with multiple, redundant sources of information. The descriptions of the seven countermeasures are excerpted from an early Semantic Hacking project paper (Cybenko et al., 2002c).

2.2 Examples of single-source countermeasures

In situations where information flows from a single source (e.g., single sensor, website, or official report), the following countermeasures were proposed.

2.2.1 Authentication and trust ratings

This method involved authenticating the information source and assessing its long-term reliability. Public Key Infrastructures and other certification tools could verify the identity of a source, while performance-based scoring could track its historical accuracy. Such systems, while conceptually simple, would require broad social or institutional consensus to become widely effective.

2.2.2 Information trajectory modeling

This approach modeled how a stream of information should logically evolve over time, allowing deviations from expected patterns to trigger suspicion. For example, weather data from a sensor could be compared to historical norms or predicted forecasts to detect anomalies, much like how financial analysts flag unusual stock price movements.

2.2.3 Ulam games

Inspired by Stanislaw Ulam's work on adversarial questioning (e.g., "20 Questions" with some deceptive answers), this model explored how to extract truth from a source known to include a limited number of falsehoods. While this approach is more applicable in dialogic or protocol-based contexts such as negotiation or stepwise authentication, it revealed the logic of interacting with a partially compromised information source. Several researchers have investigated this problem, using ideas from error-correcting codes and other areas (Mundici and Trombetta, 1997).

2.3 Multiple-source countermeasures

In modern digital ecosystems, most information comes from a mix of sources (e.g., news media, user forums, social networks, etc.). Recognizing this, the Semantic Hacking Project also proposed methods for countering attacks in environments with multiple, potentially colluding actors.

2.3.1 Collaborative filtering and reliability reporting

Borrowed from e-commerce and recommender systems, this method scored information sources based on user feedback and community consensus. While widely used in online commerce to assess vendors, its application in national security settings would involve developing similar trust layers across news sources, intelligence streams, and public discourse.

2.3.2 Byzantine generals models

Adapted from distributed computing, this model examined how to determine the reliability of actors in a group when some participants may be deceitful. Although initially designed for protocol-heavy systems, the model raised important questions about group dynamics in deception detection, especially in coalition-based environments.

2.3.3 Detection of collusion

This approach explored how automated tools could detect coordinated misinformation campaigns. For example, it could analyze message patterns in financial forums to identify multiple posts pushing the same deceptive narrative and flagging suspicious clusters or unusually aligned messaging across supposedly independent accounts.

Interestingly, the team discovered that the challenge wasn't just detecting outliers, but identifying statistical clusters that mimicked organic diversity but were actually manufactured consensus. Detecting this kind of subtle alignment remains one of the hardest problems in modern disinformation analysis.

2.3.4 Linguistic analysis

This countermeasure used stylometry and other linguistic fingerprinting techniques to determine authorship of supposedly unrelated texts. If dozens of pseudonymous posts share the same stylistic features, there's a strong chance they were authored by a single source. While less effective for very short content (like tweets), this technique remains powerful when analyzing blogs, articles, or coordinated documents.

2.4 Focused case study: the NEI Webworld pump-and-dump operation

Around the time of the Semantic Hacking Project, a particularly vivid example of a semantic (or cognitive) attack was the 1999 NEI Webworld stock manipulation case. A defunct company with little actual value became the center of a disinformation campaign when three individuals bought up shares at $0.05–$0.17 and then used over 500 deceptive messages across Internet message boards to claim the company was on the verge of a lucrative buyout.

By inventing third-party interest and repeating inflated claims through multiple pseudonymous accounts, they drove the share price to $15 within a single day and made $364,000 in profit before selling off their positions.

While technically simple, the attack exploited the perception of independent corroboration and urgency, which was exactly the kind of exploit the Semantic Hacking Project warned about. In this case, several countermeasures proposed by the project might have exposed the deception:

- Collaborative filtering and trust reporting could have lowered the credibility of new or unverified users making bold claims.
- Information trajectory modeling would have flagged the extreme and sudden stock price movement as anomalous.
- Linguistic analysis could have revealed that the hundreds of seemingly independent posts were likely authored by only three people.

The NEI case demonstrated that cognitive attacks are not theoretical. They have financial, legal, and societal impacts. And they are often carried out using nothing more than manipulation of language and identity in open forums (Cybenko et al., 2002c).

2.5 Strategic insights for today's cognitive security landscape

While the Semantic Hacking Project produced technical frameworks and early prototypes, its broader strategic insights offer enduring relevance for today's AI-powered information terrain. Three themes in particular stand out.

2.5.1 Perception can be weaponized without ever touching the facts

One of the Project's central realizations was that control over interpretation often matters more than control over content. Cognitive attacks succeed not by hiding or falsifying data, but by subtly steering how people understand it. Many of the proposed countermeasures (e.g., like information trajectory modeling or linguistic clustering) were grounded in the recognition that believable distortions are more dangerous than obvious falsehoods. This principle echoes through today's influence operations, where AI-generated content increasingly reinforces persuasive frames and synthetic consensus without needing to invent anything outright. The battleground is not information access, but meaning assignment.

2.5.2 Definitions are battlespaces

The Project foresaw the strategic importance of framing and terminology. When key terms like "sentient," "genocide," "life," or "threat" are up for debate, so too are the conclusions that follow from them. Whoever gets to define the terms often shapes the outcome. This battle over semantic framing (which some contemporary DoD staff refer to as "term warfare") is only becoming more intense in an age where digital repetition, generative AI, and social media can entrench contested definitions rapidly and globally.

2.5.3 Theory itself is a story that can be weaponized

The Project's deeper insight was that even the frameworks we use to determine what is true (our hypotheses, laws, and scientific narratives) are not immune to manipulation. Theories can be introduced, reframed, or dismissed not only through evidence, but through narrative persuasion. In today's information landscape, the contest over "truth" is also a contest over what counts as legitimate knowledge. This likely is not just a modern or future of warfare problem; it likely dates back to the earliest forms of warfare and those who knew how to wield information, law, history books, and the "truths" about the heavens were the ones winning the war without firing a shot.

2.5.3.1 From semantic hacking to AI-era cognitive security

Although the Semantic Hacking Project formally concluded in 2003, the intellectual arc it launched has continued to evolve. One major milestone came in 2014, with the AAAI Spring Symposium on "Social Hacking and Cognitive Security on the Internet and New Media," which convened experts to examine how digital narratives and online behaviors intersected with emerging information threats (AAAI, 2014). Though the term "cognitive security" had not yet fully crystallized, the concepts discussed mirrored many of the Semantic Hacking Project's insights especially regarding framing, deception, and interpretive manipulation.

Waltzman (2017) delivered congressional testimony that popularized the term "cognitive security" in national security circles (2017). The Information Professionals Association (IPA, 2025) has since grown into a practitioner hub for influence professionals across military, intelligence, academic, and commercial sectors. Meanwhile, NATO began engaging more formally with these ideas through its Innovation Hub and subsequent symposiums. Its 2021 publication, "Countering Cognitive Warfare: Awareness and Resilience," brought the term into broader use across allied military structures (NATO, 2021).

These developments reflect a growing recognition that the space the Semantic Hacking Project explored around how meaning is manipulated and truth is contested is now the central battleground for information-age conflict.

2.5.3.2 The need for cognitive testbeds

To contextualize how the insights from the Sematic Hacking Project can help with cognitive security issues today, let us go over a few open issues in the field. One of the key practitioner demands emerging today is for cognitive security to develop testbeds and proving grounds similar to those used in cyber and kinetic operations. Just as weapons are stress-tested in controlled environments, influence operations must be simulated, evaluated, and refined outside of live conflict zones (C4ISRNET, 2021).

The Semantic Hacking Project's suggestion of modeling information trajectories and simulating adversarial questioning (like Ulam Games) could provide a conceptual foundation for these test environments. By building simulations where varying definitions, belief systems, or stimuli are introduced under controlled parameters, researchers could begin developing a formal Test & Evaluation (T&E) doctrine for narrative engagement. This would help institutions gauge not only what messages are persuasive, but "why," "whom," and under what shifting contextual conditions.

2.5.3.3 The limits of deepfake detection and the need for psychophysics of belief

Much of today's discourse on AI and disinformation focuses on detecting deepfakes. But the core issue is not whether something is fake; it's whether people believe it or not. This gap between reality and perception demands a new line of research: the psychophysics of belief.

The Semantic Hacking Project already emphasized that what counts as "authentic" or "real" can shift based on context, repetition, and source credibility. To build on that, modern cognitive security must investigate which perceptual and semantic qualities make a message seem true (even when it is not) and how those thresholds change across populations and over time. Variables could range from pixel resolution and audio quality (low-level) to perceived authority, body language, or context (high-level). Regular empirical testing across demographic and cultural groups would be required to build models of what people consider trustworthy, how those judgments

evolve with exposure, and what kinds of countermeasures might restore healthy skepticism. As the Semantic Hacking Project foreshadowed, even the definition of what is "fake" or "real" is vulnerable to manipulation.

2.5.3.4 Financial and economic warfare in the cognitive age

The NEI Webworld case study foreshadowed a now-mainstream concern: information can move markets. Today's financial ecosystem, with its heavy reliance on algorithmic trading and AI-driven analysis, is even more susceptible to cognitive manipulation.

A modern systemized example would be the firm Hindenburg Research (2025) that has built a business model around investigating companies and releasing public reports which often triggers massive stock selloffs from perceived fraud or mismanagement. While this form of financial activism may serve a regulatory purpose, its success relies on controlling narrative timing and information asymmetry. On the darker end of the spectrum, scams in the cryptocurrency world (ranging from pump-and-dumps to deepfake-driven wallet theft) demonstrate how deception has become a tactical weapon.

Another notable incident occurred in 2023, when an AI-generated image of an explosion at the Pentagon caused a 15-min dip in the U.S. stock market (New York Times, 2023). That brief but measurable loss revealed how much economic stability now hinges on real-time narrative control. Just as troubling, adversaries may not need to hack infrastructure directly, they can attack trust in the systems that interpret it. The Semantic Hacking Project's emphasis on collusion detection, trajectory modeling, and information source reliability could provide valuable tools for countering these threats in finance.

3 Conclusion: revisiting the past to secure the future

This paper was written not just to highlight an early and underappreciated effort in the study of cognitive security, but to challenge the assumption that today's problems are wholly new. In the face of fast-moving developments in cognitive security and AI, there is a growing tendency to believe that novel threats require entirely novel solutions. But as the Semantic Hacking Project showed nearly 25 years ago, many of the core dynamics of cognitive conflict were already understood, modeled, and even partially addressed.

We are fortunate to have one of the project's principal contributors as a co-author, able to preserve and share insights that could easily have been lost. Unfortunately, this kind of historical continuity is rare in defense and national security research. Because so much work is classified, ephemeral, or siloed in closed communities, institutional knowledge often disappears when individuals retire, change roles, or move on. The result is a field where rediscovery often replaces progression, and where the wheel is reinvented more often than we like to admit.

Cognitive security is too important, too urgent for that kind of amnesia. If we are to build effective doctrines, testbeds, and research programs, we must begin by mapping what has already been tried, what was learned, and what still remains unsolved. The Semantic Hacking Project deserves a place in that lineage not as a historical footnote, but as a foundational effort whose questions are still guiding us, even if the technologies around them have changed.

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

PT: Writing – original draft. SG: Writing – original draft.

Funding

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

Conflict of interest

SG was employed at MAD Warfare.

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

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

References

AAAI. (2014). Social hacking and cognitive security on the internet and new media. In: Reports of the 2014 AAAI spring symposium. Available online at: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2550 (Accessed July 16, 2025).

C4ISRNET. (2021). Protecting people from disinformation requires a cognitive security proving ground. Available online at: https://www.c4isrnet.com/opinion/2021/02/10/protecting-people-from-disinformation-requires-a-cognitive-security-proving-ground/ (Accessed July 16, 2025).

Cybenko, G., Giani, A., Heckman, C., and Thompson, P. (2002c). Cognitive hacking: technological and legal issues law tech 2002 November 7–9, 2002. Cambridge, Massachusetts: Acta Press.

Cybenko, G., Giani, A., and Thompson, P. (2002a). Cognitive hacking and the value of information. Workshop on economics and information security, May 16–17, 2002, Berkeley.

Cybenko, G., Giani, A., and Thompson, P. (2002b). Cognitive hacking: a battle for the mind. *IEEE Comput.* 35, 50–56. doi: 10.1109/MC.2002.1023788

Cybenko, G., Giani, A., and Thompson, P. (2004). "Cognitive hacking" in Advances in computers. ed. M. Zelkowitz, vol. 60, 36–75.

Hindenburg Research. (2025). Available online at: https://hindenburgresearch.com/ (Accessed July 16, 2025).

IPA. (2025). Information professionals association. Available online at: https://information-professionals.org (Accessed July 16, 2025).

Libicki, M. (1994). The mesh and the net: speculations on armed conflict in an age of free silicon. National Defense University McNair paper 28. Available online at: http://www.ndu.edu/ndu/inss/macnair/mcnair28/m028cont.html (Accessed July 16, 2025).

Mundici, D., and Trombetta, A. (1997). Optimal comparison strategies in Ulam's searching game with two errors. *Theor. Comput. Sci.* 182, 217–232. doi: 10.1016/S0304-3975(97)00030-3

NATO. (2021). Countering cognitive warfare: awareness and resilience. NATO review. Available online at: https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html (Accessed July 16, 2025).

New York Times (2023). An A.I.-generated spoof rattles the markets. Available online at: https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html (Accessed July 16, 2025).

Thompson, P. (2003). Utility-theoretic information retrieval, cognitive hacking, and intelligence and security informatics. In Proceedings of the joint conference on information sciences, 26–30 September, 2003, Cary, North Carolina, p. 452–457.

Waltzman, R. (2017). The Weapoonization of information: the need for cognitive security, expert insights published 27 April 2017. Testimony presented before the Senate Armed Services Committee on April 27, 2017.

Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact