RETHINKING ANTI-MISINFORMATION AI

Vidya Sujaya[1, 3], Kellin Pellrine[1, 3], Andreea Musulan[2, 3] & Reihaneh Rabbany[1, 3]
[1] McGill University [2] Université de Montréal [3] Mila

ABSTRACT

This paper takes a position on how anti-misinformation AI works should be developed for the online misinformation context. We observe that the current literature is dominated by works that produce more information for users to process and that this function faces various challenges in bringing meaningful effects to reality. We use anti-misinformation insights from other domains to suggest a redirection of the existing line of work and identify two opportunities AI can facilitate exploring.

1 INTRODUCTION

AI-based proposals fighting online misinformation are dominated by works that focus on producing more information about existing information artifacts (eg. whether a piece of content is fake news (Khanam et al., 2021), or whether some online social media user is authentic ((Masood et al., 2019))). However, as more people flood online spaces, and the popularization of LLMs as consumer products like ChatGPT and Grok reduces language and time limitations in producing online content, we question whether this focus yields the most meaningful solutions. So, we look at anti-misinformation literature from other domains to outline the limits of current anti-misinformation AI works and identify opportunities for more meaningful AI-based solutions. We define misinformation as information that cannot be supported by factual evidence from a reputable source but leave the definition of 'reputable' (and related phrases like 'good quality information') beyond the scope of this work. We proceed with trusting the readers' understanding of them and suppose a shared definition of what constitutes quality information for the rest of this paper.

2 ARTIFACT-BASED ANTI-MISINFORMATION PROPOSALS

If we picture social media, its users, and the content within it as an 'online information ecosystem', we can think of the first as the infrastructure, and the latter two as artifacts within it. Then, if a work heavily relies on using or producing information about artifacts, we can refer to them as 'artifact-based'. The popular category of AI anti-misinformation research dedicated to misinformation detection clearly falls under this description as it relies on analyzing information within content (Islam et al., 2020), or of users (eg, posting behavior, following, and followers) to determine some measure of a characteristic of the artifact (eg, veracity, authenticity) (Shu et al., 2019). The same can be argued for related work such as information verification, automated labeling/explanation generation, and explorations of creating and distinguishing AI-generated misinformation (Zhou et al., 2023). Also, although not as straightforward, we argue that recommendation and ranking systems are artifact-based, as they depend on information within the content or the users in determining the results (Wang et al., 2022; Sallami et al., 2023). Finally, we also consider simulated works to be artifact-based, as they focus on studying the movement of, relationships, and effects between artifacts (eg. Yilmaz & Ulusoy (2022) simulate the propagation of misinformation within an online social network, whilst Touzel et al. (2024) create a simulation of a group of LLM-based agents, and test the effects of manipulation on the agents on election results within the group). In contrast, under this categorization, a non-artifact-based work focuses on the infrastructure. This can be exemplified by a work, that for example, explores how a video-sharing platform's content sharing function effects information propagation, compared to sharing functions of a micro-blogging platform (ie. sharing links, compared to the ability to 'repost' and 'quote' respectively). With that, a distinctive characteristic between the two categories follows: artifact-based works have information about artifacts, or the artifacts themselves, imposing an effect onto the work, whilst infrastructure-based works impose an effect on all artifacts.

As most anti-misinformation AI literature then falls under the artifact-based description, we question how meaningful they are in fighting the general misinformation problem. One clear bottleneck of the existing proposals is their reliance on the parties that facilitate the infrastructures to implement some action based on their results (eg. using results of detection algorithms for content/user moderation, ranking, and recommendation). The relevance of this bottleneck is highlighted by the recent change in Meta's moderation program in the United States, shifting from fact-checking to a community notes system (as reported in Reuters (2025)). We acknowledge the challenges and complexities of realizing such steps, and leave it for a different discussion. However, if we ignore this possibility of directly enforcing what artifacts are accessible or able to be on the information landscape, the value of these artifact-based proposals is rendered down to the ability to produce more information about artifacts. Their value is then determined by how many people choose to access, process, and use this information, and how meaningfully they do it. Additionally, we note that many non-AI Computer-Science-based interventions still fall under the artifact-based description, and therefore follow the above effects. As illustrative examples, proposals of plugins and platforms for crowd-sourcing human-judged content quality (Jahanbakhsh & Karger, 2024), and usage of tamper-proof blockchain to track content propagation and credibility Seneviratne (2022), functions to produce information on the content artifact. Now, our question becomes how likely this is. We present our answer in the following section, looking at several insights from other domains.

3 Pitfalls of the Artifact-Based Approach

First, acknowledging the declining trend of human attention span (Mark, 2023) and how much existing infrastructure operates on an attention economy, we are unsure of the feasibility of demanding more attention from individuals toward extra information. Further, we know from psychological perspectives of biases of the human mind that show how we are selective of what content we consume and choose to accept, and perhaps not in a way that makes quality information the most appealing. This includes the attraction of our minds to content that provokes negative emotions (Acerbi, 2019), and consideration of alignment with preexisting intuition and in or out group-based measures of credibility when accepting information (Ecker et al., 2022). Second, even if we can successfully make quality information reach and be consumed by individuals, we also know of pitfalls like the continued influence effect (CIE) (where users' beliefs may be corrected, but their actions are still based upon their previous uncorrected ones (Ecker et al., 2022)), which render efforts of misinformation correction less meaningful in domains where the effect on human action is valuable (eg. ensuring people make healthcare choices based on factual medical information, or vote during elections without getting affected by rumors or conspiracies). These two points are not comprehensive of the findings of all domains regarding misinformation, but outline the insufficiency of the main function of current approaches in producing more information.

Since the human mind and behaviors are part of the challenge, perhaps we need to teach individuals to be better at their information practices. In such a situation, one might like to hand over the task of spreading how to meaningfully use quality information to a different third party, namely those within media literacy and education domains. However, conflicting perspectives on the utility and role of these domains' approaches to misinformation exist. For example, Bulger & Davison (2018) points out how media literacy's impact is dependent on context, can have very minimal impact in some, or result in overconfidence (which Lyons et al. (2021) suggests as an important variable in studying online spread of low-quality information). Then, in 2010, Nyhan & Reifler (2010) found corrective approaches of misinformation to carry the risk of the backfire effect (increased misperception). More recently, a follow-up study finds corrective approaches to still have a positive effect, albeit decaying over time (Nyhan, 2021). There are existing works that rethink approaches to media literacy (eg. Mihailidis & Viotty (2017) suggests media literacy to move beyond building people's information consumption skills, towards caring, and considering civic impact). For now, these points uncover the challenges within media literacy as a domain, and lead us to the same stance of how it is insufficient for artifact-based works to simply produce or distinguish 'quality' information, and rely on the consumers for meaningful use.

4 REDIRECTIONS

So, how should we move forward? We think that exploration should go beyond artifact-based proposals, and that existing artifact-based works be made more meaningful by considering learnings and possible perspectives of other domains. We expand on both points in the next subsections.

4.1 MAXIMIZING ARTIFACT-BASED PROPOSALS

Rethinking Evaluation: The Continued Influence Effect (CIE) pitfall raises the question of what we are trying to achieve when we say anti-misinformation: do we care only about what information people accept, or how that information is used too? This sets up one example of how to improve existing artifact-based works, that is, go beyond the current goals of evaluating how effective proposals are in correcting misinformation in the individual's minds (Mark, 2023), and rethink what to measure by asking what effect the proposal's product should realize. Further, earlier points on cultural and contextual dependencies of the effects of misinformation should also be considered in metric measurement, and translated to the kinds of data collected and used in experiments. Long-term measurement of impact (eg. checking for the persistence of information correction) may also be necessary as we see evidence of how positive effects of correction may decay over time (Nyhan, 2021). In this direction, an example action point can be a long-term test of label or explanation generation tools. This involves testing the effects of such tools over time, using questionnaires sent at different time periods following a user's first encounter with the generated label or explanation. The questionnaires could test for the persistence of misinformation correction (if any), and check for what understanding of the information is used when users are prompted to make some decision.

Improving assumptions of the human: This links to our earlier point of existing pitfalls and biases of the human mind but also recent literature questioning the human relationship with misinformation. This includes how human information consumption habits aren't always rational (Munn, 2024), how misinformation's effects and risk levels are dependent on factors like cultural contexts and political views (Sample et al., 2018; Rampersad & Althiyabi, 2020; Tokita et al., 2024), and complexities translating knowledge about the relationship between people and misinformation between online and offline spaces (Kozyreva et al., 2020; Altay et al., 2023), and between different contexts [1]. In this direction, an example action point can delve into the design of agents in works that involve simulations. In works like Touzel et al. (2024), where the exploration relies on an AI-based simulation of individuals (here, used to explore the effects of manipulation on the simulated agents), a more comprehensive assumption on the human model can include designing for different degrees of rationality, and emotional factors, rather than stopping at 'human-like' personas.

m

4.1.1 BEYOND ARTIFACTS

Moving on, we highlight two opportunities for anti-misinformation work that provide an alternative to the current set of artifact-based proposals.

Studying the medium: We find an opportunity for exploration beyond artifacts to move to the other component of the information landscape, that is, the infrastructure. Intuitively, think of the different functions users are given in video-exclusive platforms, compared to microblogs (eg. amplifying content on one can be done with a repost or quote function, while the other relies on sharing links through other platforms). Further, think of differences between long-form video exploration mechanisms with short ones (eg. after a video ends, are you presented with a list to choose from, or are you scrolling to the next one)? The idea that the infrastructure or medium matters aligns with the communication theory 'the medium is the message' (McLuhan, 2019), and starting medium designs to explore can be informed by works like the Nudge Deck (Konstantinou & Karapanos, 2023). While it may be difficult to experiment with design changes of the medium in real life, AI-facilitated simulations, if made more robust and aligned with human belief systems, can be a sandbox for experiments centered on the medium. Initial experiments can simulate the spread and effects of some misinformation seed in the presence of two different social media platforms, one with a different

[1] Kozyreva et al. (2020) explains the distinctions of online and offline environments, and how each impact people's behaviors, and Altay et al. (2023) shares the need to better study relationship between online information practices of people, their true beliefs, and their action taking process

feature compared to the other (eg. sharing capabilities limited only to reposting and tagging, as opposed to reposting, tagging and quoting). Each feature would presumably allow users to practice different behaviors, depending on what is afforded to them by the feature. Here, 'affordance' follows the definition in Davis (2023), that is how features of a technology shape but do not determine its functions and effects.

'Guardian AI': This second opportunity builds off of previous 'redirections' and revisits the idea of a Guardian AI. The idea has previously been described as functioning as an automated recommendation filtering (Rumbelow, 2022). But, with better goals and knowledge of metrics, the human model, and an understanding of what would constitute healthy mediums, perhaps the Guardian AI can consolidate existing artifact-based tools into a comprehensive end-user tool that accompanies users navigating the online information landscape. Knowledge about mediums can make the comprehensive tool considerate of the human-mind-related challenges introduced earlier (ie. attention span, CIE, overconfidence backfire effect). This can come in the form of how the Guardian AI is designed, but also a warning function against online platform features that may nudge the user against mindful information processes, as the user goes through different platforms. Further, recognizing the importance of contexts (eg. information topic, user beliefs, user culture) when interacting in the online information ecosystem, such a tool can balance between personalization (to user goals for example), and shared values. Of course, such a comprehensive end-user tool would be a challenge to design, and details on the constitution of a 'healthy' Guardian AI would require collaboration from different domains, and perhaps civic engagement for reviews.

5 Limitations

We observe three main limitations of our work. First, our claim of the value of artifact-based approaches being limited to extra information production is made with the assumption that options like moderation, which requires third-party implementation, are unfeasible. This needs to be re-evaluated otherwise. Second, our exploration of other domains' perspectives is still limited mostly to media literacy and psychology. Works in other domains (eg. anthropology, sociology) still need to be consolidated. This also includes consolidating non-AI Computer Science approaches against misinformation. Finally, our suggested redirections require the adoption of something new (eg. a medium or end-user tool), and an agreement on what constitutes healthy information practices and 'quality' information, which may be challenging. However, a new research-backed platform informed by the study of mediums can afford us to aim for goals that are difficult to achieve when focusing only on artifacts. For example, a platform can be designed for a more belief-neutral goal of caring for people's attention span, while the same goal may be difficult to set in an artifact-only approach. The agreement challenge is connected to how a Guardian AI should be calibrated between personalization, and fixed goals or design features.

6 Conclusion

In this position paper, we introduced the artifact-based concept and outlined how most AI-based anti-misinformation literature falls under the category. Then, referring to insights on, but not limited to, pitfalls like CIE, and challenges with media literacy, we describe how artifact-based tools' function of producing extra information is insufficient in meaningfully combating misinformation. We are not suggesting an abandonment of artifact-based proposals, but the need to allocate resources to answering other questions. Existing artifact-based research can be redirected to be more meaningful, by, for example, considering metrics based on desired real-world impact. Moving forward, the ability to conduct robust simulations, unique to AI capabilities, can in turn help explore new questions on medium design, and prospects of a comprehensive end-user tool that accompanies individuals as they navigate the online information ecosystem.

References

Alberto Acerbi. Cognitive attraction and online misinformation. *Palgrave Communications*, 5(1), 2019.

Sacha Altay, Manon Berriche, and Alberto Acerbi. Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ society*, 9(1):20563051221150412, 2023.

Monica Bulger and Patrick Davison. The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education*, 10(1):1–21, 2018.

Jenny L Davis. 'affordances' for machine learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 324–332, 2023.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1 (1):13–29, 2022.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82, 2020.

Farnaz Jahanbakhsh and David R Karger. A browser extension for in-place signaling and assessment of misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2024.

Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, pp. 012040. IOP Publishing, 2021.

Loukas Konstantinou and Evangelos Karapanos. Nudging for online misinformation: a design inquiry. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, pp. 69–75, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701290. doi: 10.1145/3584931.3607015. URL https://doi.org/10.1145/3584931.3607015.

Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21 (3):103–156, 2020.

Benjamin A Lyons, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan, and Jason Reifler. Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23):e2019527118, 2021.

Gloria Mark. *Attention span: A groundbreaking way to restore balance, happiness and productivity*. Harlequin, 2023.

Faiza Masood, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani, and Mansour Zuair. Spammer detection and fake user identification on social networks. *IEEE Access*, 7:68140–68152, 2019.

Marshall McLuhan. The medium is the message (1964). In *Crime and media*, pp. 20–31. Routledge, 2019.

Paul Mihailidis and Samantha Viotty. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American behavioral scientist*, 61(4): 441–454, 2017.

Luke Munn. Misinformation's missing human. *Media, Culture & Society*, 46(6):1287–1298, 2024. doi: 10.1177/01634437241249164. URL https://doi.org/10.1177/01634437241249164.

Brendan Nyhan. Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15):e1912440117, 2021.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

Giselle Rampersad and Turki Althiyabi. Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, 17(1):1–11, 2020.

Thomson Reuters. Non-cooperative games. *Meta to end fact-checking program on Facebook, Instagram in U.S.*, 2025. URL https://www.cbc.ca/news/business/meta-fact-checking-program-ends-1.7424631.

Jessica Rumbelow. Guardian ai (misaligned systems are all around us.), 2022. URL https://www.lesswrong.com/posts/iHLJtbdFwsoNWZg3e/guardian-ai-misaligned-systems-are-all-around-us.

Dorsaf Sallami, Rim Ben Salem, and Esma Aïmeur. Trust-based recommender system for fake news mitigation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pp. 104–109, 2023.

Char Sample, John McAlaney, Jonathan Z Bakdash, and Helen Thackray. A cultural exploration of social media manipulators. *Journal of Information Warfare*, 17(4):56–71, 2018.

Oshani Seneviratne. Blockchain for social good: Combating misinformation on the web with ai and blockchain. In *Proceedings of the 14th ACM Web Science Conference 2022*, pp. 435–442, 2022.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 436–439, 2019.

Christopher K Tokita, Kevin Aslett, William P Godel, Zeve Sanderson, Joshua A Tucker, Jonathan Nagler, Nathaniel Persily, and Richard Bonneau. Measuring receptivity to misinformation at scale on a social media platform. *PNAS nexus*, 3(10):pgae396, 2024.

Maximilian Puelma Touzel, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, et al. A simulation system towards solving societal-scale manipulation. *arXiv preprint arXiv:2410.13915*, 2024.

Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 3673–3684, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512263. URL https://doi.org/10.1145/3485447.3512263.

Tolga Yilmaz and Özgür Ulusoy. Misinformation propagation in online social networks: game theoretic and reinforcement learning approaches. *IEEE Transactions on Computational Social Systems*, 10(6):3321–3332, 2022.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581318. URL https://doi.org/10.1145/3544548.3581318.

A Appendix

You may include other additional sections here.