RETHINKING ANTI-MISINFORMATION AI

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

ABSTRACT

This paper takes a position on how anti-misinformation AI 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 insights from other domains to suggest a redirection of this body of works.

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 contents (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 works such as information verification, automated labeling/explanation generation, and explorations of creating and distinguishing AI-generated misinformation (Zhou et al., 2023). Also, though not as straightforward, we argue that recommendation and ranking systems are artifact-based, as they depend on information within content or of users in determining 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).

As most anti-misinformation AI literature then falls under the artifact-based description, we question how meaningful they are in fighting the misinformation problem. One 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). 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 affecting what artifacts are made accessible on the information landscape, the value of artifact-based proposals render down to their ability to produce more information about artifacts. Their effect is then determined by how many people choose to access, process, and use this information, and how meaningfully they do it. Now, our question becomes how likely this is. In the next paragraph, we briefly look at other domains to answer this question.

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 elicits negative emotions (Acerbi, 2019), and consideration of alignment with preexisting intuition and group-based or inactive credibility measures 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. One might like to hand over the task of meaningfully spreading and using quality information to a different third party, namely those within media literacy and education domains. However, conflicting perspectives on the utility and role of common approaches towards media literacy Bulger & Davison (2018); Lyons et al. (2021); Nyhan (2021) lead us to the same insufficiency stance.

3 REDIRECTIONS

So, how should we proceed? First, we think that exploration should go beyond artifact-based proposals, and find an opportunity for exploration to the information landscape itself, namely the infrastructure. Intuitively, think of differences between long-form video exploration mechanisms and 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 significance of the idea that infrastructure matters aligns with communication theory, 'the medium is the message' (McLuhan, 2019). Starting works can explore what design choices (Konstantinou & Karapanos, 2023) afford Davis (2023) users in their information consumption practice. A simple experiment can test how different content-sharing functions (eg. reposting versus quoting) affect the time a user spends engaging with the content, and how much of it is retained over time. While such an experiment may be difficult and expensive 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.

Second, existing artifact-based proposals should aim for meaningful real world effects, which can be done by considering knowledge perspectives of other domains. Some examples include: **(1) Rethinking metrics:** The 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. **(2) Improving assumptions of the human:** This links to our earlier point of existing pitfalls and biases of the human mind, changing attention spans, and 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 effect may be limited to peripheries, and complexities translating knowledge about misinformation from online to offline, and between different contexts. While the rationality point can simply inform how individuals in simulations are designed to interact with content, the less-answered questions point towards the need for collaboration between domains to handle misinformation's complexities.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

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

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.

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.

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.

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.

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.

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.

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.