The role of AI in the misinformation ecosystem: virtuous or vicious?

Claudia b Claros[1], Kimiz Dalkir[2]

[1] SIS, McGill University, 27 3661 rue Peel, Montréal, CA
[2] SIS, McGill University, 27 3661 rue Peel, Montréal, CA

Abstract

Artificial Intelligence (AI) systems have become ubiquitous and often invisible actors in Information and Communication Technologies (ICTs). The complexity (and intelligence) of these systems varies, but we end-users can't easily differentiate them from human actors any longer. While AI systems allow us to respond more efficiently to misinformation, they also present new challenges: they can create, spread, or suppress (mis)information. We discuss some of the immediate, medium-term, and long-term challenges of AI to the spread of misinformation and discuss safety in terms of in-system and cross-system strategies, concluding that AI integration into ICTs doesn't automatically resolve problems arising from misinformation spread. In response to cyclical information disorders, ecosystems might undergo a process of reintermediation between information-seekers and trusted providers of knowledge and resources, whether human or machine.

Keywords

misinformation, disinformation, AI, information ecosystems, reintermediation.

1. Introduction

The technological progress of the last decade has disintermediated information ecosystems. Traditional intermediaries to information and resources are replaced by access to the web and social media, facilitating the spread of (mis)information through weak ties in digital, and therefore analog, social networks. As a result, users' need to critically interpret information alongside the cognitive load involved in navigating information environments has exponentially increased. As AI and other intelligent systems become explicit participants in knowledge creation or invisible actors within information structures, we might now observe a process of digital re-intermediation as end-users seek out trusted sources, whether human or artificial, to lighten the cognitive load of daily verification, evaluation and decision making tied to information seeking and use. In this context, AI systems risk becoming vectors of misinformation consumption and spread. We discuss the immediate, medium-term, and long-term impact of AI systems to the spread of misinformation, approaches to safety, and coordinated, ethically 'virtuous' integrations of AI.

2. Misinformation & AI: Immediate, medium-term, and long-term challenges

Misinformation online can be defined as meaning encoded in a specific format (text, audio, audio-visual media) that inaccurately represents that which it portrays as verified by conventional standards of evidentiary support, typically the scientific consensus (Lewandowsky, 2020). Misinformation has certain characteristics that make it particularly cognitively attractive to us, notably negative affect, specific themes of human interest, and eliciting surprise or disgust (Acerbi, 2019; Bessi et al., 2015; Vosoughi et al., 2018). The cognitive load involved in navigating day-to-day informational environments means that most users heavily rely on priors and heuristics to make credibility evaluations (Islam et al., 2020). This dependency on our prior beliefs or 'hunches' to evaluate the veracity of information means that we are all equally susceptible to the misinformation that best conforms to our biases. In the name of efficiency and accounting for the fallibility of human judgement, AI solutions such as algorithmic content suppression have become the norm (Dalkir & Katz, 2020; Rubin, 2022).

AiOfAi'23: 3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Macao, China

Claudia.baptistaclaros@mail.mcgill.ca (C. b Claros); kimiz.dalkir@mcgill.ca (K. Dalkir)

0009-0004-4595-1763 (C. b Claros); 0000-0003-3120-6127 (K. Dalkir)

© 2023 Copyright for this paper by its authors. The use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3593/paper6.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Artificial Intelligences (AI) are complex software-based systems developed with a range of techniques that can produce outputs or make decisions that influence the environments they are in (European Union AI Act, 2021). Recent iterations of AI, such as generative AI and LLMs, can learn autonomously and produce seemingly coherent and accurate summaries of complex issues in seconds. The impact of 'intelligent' technologies making choices in our information ecosystems, including which information should be eliminated from these environments, has both immediate and long-term consequences for society. In terms of immediate challenges, AIs are only as smart as their training, are subject to manipulation, can produce misinformation, and reproduce epistemic and ethical values encoded in data and systems; as well as their corresponding ills, be it racism, sexism or other. In the medium term, these technologies are likely to cascade social, economic and political changes through society. In the long term, the potential for AI to become autonomous actors raises existential questions about human and robot ethics.

2.1. Immediate

Defining and understanding 'intelligence' in terms of these complex systems is an immediate challenge. Large language models, or LLMs, can process massive amounts of unstructured data and seemingly use natural languages coherently. However, recent attempts to benchmark LLM performance show that their accuracy at resolving even simple mathematical problems is not guaranteed (https://benchmarks.llmonitor.com/sally). Their complexity makes them unpredictable, and their accuracy varies significantly through user interaction and across datasets, languages, and models.

How the ambiguity, uncertainty and context dependent meaningfulness of natural languages is encoded into machine language remains an open scientific question (Birhane, 2023). But their apparent linguistic fluency can trick users' perceptions of their intelligence. Natural languages emerge from and are enacted through intersubjective coordination. The meaning of a sentence is context dependent, socially embedded and (mis)interpretable. In contrast, LLM-generated informational outputs are based on computational models of natural languages, trained on massive pre-existing and 'unfathomable' datasets, often created by trawling content from the web (Kaddour et al., 2023). Their accuracy and performance are relative to the amount of machine-readable data available in a specific language, for example. While a systems' accuracy might be improved by probabilistically accounting for the ambiguity of natural languages, they have (as of now) limited access to the socially embedded meaningfulness of natural language propositions.

The differences between machine and natural languages mean that evaluating an AI's informational output as true or false, or automatizing processes of information verification and evaluation, still require human intelligence at the helm. While we can automatize misinformation recognition, verifying how far a statement corresponds to the scientific or expert consensus requires the interpretive work of human experts or fact-checkers (Dalkir, 2021).

The differences between machine intelligence and our intuitive assumptions about intelligence are still unclear (Newfield, 2023). The potential for misuse based on users' misunderstanding of machine intelligence is immediate, even when intelligent systems point out their limitations. ChatGPT for example, warns users about its potential to produce misinformation ("ChatGPT may produce inaccurate information about people, places, or facts"). But these warnings are only as good as users' heeding of them, and their capability, time, and desire to independently evaluate the AI generated content. Demands on users' information literacy skills are likely to increase as the technologies become more complex and intelligent. Critical AI literacy skills include users' understanding of the way their prompts and queries determine the systems' response. How educators should account for AI generated information in their classrooms, as well as their use in plagiarism and fraud, are urgent and open questions of immediate relevance, which are currently handled ad hoc.

2.2. Medium-term

In the medium term, AI provides us with countermeasures against information disorders while simultaneously catalyzing new ones. It can simulate the spread of misinformation through social networks and track false narratives even in the absence of their original source (Vosoughi et al., 2018; Aïmeur et al., 2023). Meanwhile, it can accelerate the technical capabilities of bad actors to produce and spread disinformation (Kertysova, 2018). An information race in which new AI is developed to counteract or compete with other AI is an emerging cycle within our information ecosystems with far ranging geopolitical, economic, and social implications.

Misinformation and the AI solutions rolled out to mitigate it, such as algorithmic content suppression, continue to raise existential questions on democratic legitimation (Lance & Livingston, 2018). Citizens need accurate, timely, and up-to-date information to make informed political choices, and while AI can 'clean' information ecosystems, its accuracy on headline issues for which it might not have sufficient data is still questionable. The use of proprietary and black-boxed AI in the moderation of misinformation and User Generated Content (UGC) is now embedded into social media platforms. Disowning traditional information sources (i.e. the press and journalistic media), information technology corporations have near complete ownership of and oversight over platforms for public information access. This gives them a substantial capability to influence narratives about political and economic seats of power (Jaeger & Burnett, 2010).

The most influential or efficient AI are likely to grant uncommon advantages to entities (nations, corporations, individuals) that can use them in their favor, increasing international competition and creating an environment in which not competing is not an option (Ramonet, 1998). The reality of creating and developing AI technologies within a profit and

competition driven market paradigm means that workers with few legal or social protections might experience heightened precarity as efficiency is prioritized and innovation disrupts industries (Cressman, 2019). The integration of AI into society simultaneously resolves and creates new 'socio-technical gaps' between the technology and how it interacts with the social worlds of users, making specific upcoming challenges hard to predict and address (Dobbe et al., 2021).

2.3. Long-term

In the long term, some argue that rapidly evolving generations of AI will be determined by evolutionary imperatives such as competition (Hendrycks, 2023). They might become autonomous agents that are independent from human input and act to achieve goals, and whether these align with human interests or not, self-preservation to ensure they achieve their goals might itself become a goal. Other views suggest that intelligent actors, human and artificial, can coexist in coordinated and virtuous information ecosystems in which the strengths of each intelligence will contribute to collective sensemaking and knowledge creation (Friston et al., 2022). To this more utopian end, we identify the need for coordinated within and cross-system approaches to AI safety.

3. Virtuous AI: safety within and across systems

AIs are now actors in our information ecosystems. Their safe integration is defined as an emergent property of the interaction between the AI, users, and society: a socio-technical challenge that cuts across disciplines and paradigms (Dobbe et al., 2021). The key challenge of safety is that the complex systems from which artificial intelligence emerges might not necessarily align how they achieve goals with the often uncertain, contradictory and context dependent ethical values of human actors (Bengio, 2023). AI might develop concretely harmful strategies to reach a generally beneficial goal.

In view of these risks, Bengio (2023) recommends bans on systems that can act in the world ('executives'), as opposed to 'scientists', who investigate the world. Constraining immediate, medium, and long-term challenges of AI integrations into information ecosystems requires a 'yes, and?' approach that coordinates accountability checks within and around AI systems, through international regulatory frameworks and protocols; as well as safety, transparency, and quality standards (Rakoba & Dobbe, 2023). The ubiquity of these systems, their current black-boxed implementations and protection through patent and IP laws, raises questions on how far these strategies will be realistically implemented.

3.1. In-system: normative optimization

Safety mechanisms can be designed and encoded within AI systems. In terms of misinformation, machine learning strategies can quickly improve the accuracy of scientist AIs. For example, the same technologies used to attack AIs and make them produce misinformation can be used to train them to recognize misinformation (Amri et al., 2022). However, Laufer et al. (2023) show that optimization can implicitly introduce normative assumptions into the system. Benjamin (2020) identifies multiple dimensions by which new technologies and the paradigm under which they are developed engineer and reinforce inequity.

In 2018, algorithmic recruitment tools and chatbots already reproduced bias, whether ethnic, racial, or gendered. Even when direct datapoints about race or gender were eliminated, the algorithms surmised from the rest to reproduce societal bias. Information produced by AI mirrors and validates implicit values in its design and training data. AI might improve our technical ability to identify and suppress misinformation, but by implicitly reproducing implicit and invisible epistemologies that underpin contemporary inequality and societal hierarchies, it might also reinforce the social conditions that lead people towards conspiracies and radicalization.

Floridi (2023) proposes a 'value double-charged' thesis, where the ethical ramifications of technology are evaluated through the different vectors that generate, contextualize, and apply the technology. This suggests that the balance of forces, how the tech is designed, legislated, and used, creates an ethically neutral or virtuous equilibrium; one where the harm caused as it acts on society is considered neutral or positive. In terms of AI, these ethical calculations are now under negotiation.

3.2. Cross-system: coordinating approaches

Approaches to AI safety will likely require coordinating technical, legislative, educational and protocol-based strategies, in addition to direct bans. Currently, regulatory oversight on AI development is under discussion and will determine the legal obligations of companies to encode and implement safety measures. The EU's Artificial Intelligence Act is the first of its kind to offer a regulatory framework to mitigate risk. They classify risk in terms of unacceptable, high, limited, and minimal or no risk. Unacceptable risk is described as posing a threat to "safety, livelihoods and rights of people", such as social scoring by governments. High-risk AI systems include a wide range of technologies, active across sectors such as critical infrastructure, essential public and private services, migration, asylum, border control management, and administration of justice and democratic processes. For these systems explicit safety measures ranging from complete ban to obligations regarding risk-assessment, delineation of use, traceability of results, dataset quality, robustness, security, and accuracy will be applied. Before these technologies can be put to market, they will be authorized to do so by a judicial or other independent

body. They note the need to future-proof legislation so it can adapt to the accelerated rate of development.

Evidence for these complex systems' behavior is increasingly crowdsourced. Projects such as the Taxonomy of AI Vulnerability by the AI Risk and Vulnerability Alliance (AVID) collect instances of security related vulnerabilities or unintentional failures (https://avidml.org/), to support auditors looking to assess the risk of AI or developers who seek to build a system considering known risks. As previously mentioned, others attempt to benchmark AI accuracy and performance (https://benchmarks.llmonitor.com/sally). Initiatives seeking to help the public understand these new technologies and dispel misinformation about AI itself are also of relevance (https://www.aimyths.org/).

The importance of multidisciplinary research approaches to understand the interaction between users and technology can't be understated. Conceptual frameworks that integrate user behavior and human interests into AI design, such as Hard Choices in Artificial Intelligence (HCAI) or Human Centered Explainable AI (HCXAI) describe protocols that commit AI systems to accountability and transparency standards. Research centers such as the Distributed AI Research Institute (DAIR) and Center for AI Safety (CAIS) acknowledge the need to coordinate teams across disciplines and bridge paradigmatic chasms. Initiatives that connect policymakers, technologists, journalists, and researchers, such as Data & Society and HKW Misinformation Review, communicate scientific understandings of these technological developments and their impact to professional audiences.

4. Cyclical information disorders and re-intermediation

The relationship between humans and machines is likely to develop its own idiosyncratic languages and literacies. User interaction with LLMs already shows curious and difficult to explain phenomena such as system hallucinations or 'prompt hacking'. As we negotiate the integration of AI within information ecosystems, the specific ways they support or impede the spread of misinformation are unfolding. We can count on both, simultaneously.

Protecting the uncertainty and ambiguity of many dissenting voices, a "diversity of biases", might prove to be a necessary contradiction of social resilience to cyclical information disorders (Benjamin, 2020). Simultaneously, differentiating human generated content from AI generated content might be an equally important element of protecting ecosystems from manipulation by disinformation.

Through the coming decade, we might see a process of distributed re-intermediation, where users of the web turn to experts, information professionals and scientific communicators to evaluate new information, human or artificial remains to be seen. Further research on the impact of AI on the work of fact-checkers, journalists and other stakeholders should aim to document incoming shifts.

References

[1] Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*(1). https://doi.org/10.1057/s41599-019-0224-y

[2] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 30.

[3] Amri, S., Sallami, D., & Aïmeur, E. (2022). EXMULF: An Explainable Multimodal Content-Based Fake News Detection System. In Foundations and Practice of Security: 14th International Symposium, FPS 2021, Paris, France, December 7–10, 2021, Revised Selected Papers (pp. 177-187). Cham: Springer International Publishing.

[4] Bengio, J. (2023, 7 May). AI Scientists : Safe and Unsafe AI? Retrieved June, 25, 2023, from https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/

[5] Benjamin, R. (2020). Race after technology: Abolitionist tools for the new Jim code.

[6] Bergmann, E. (2018). Conspiracy & populism: The politics of misinformation. Cham: Springer International Publishing.

[7] Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PLOS ONE*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

[8] Cressman, D. (2019). Disruptive innovation and the idea of technology. Novation: Critical Studies of Innovation, I, 23–23.

[9] Dalkir, K. (2021). Fake News and AI: Fighting Fire with Fire?, AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA

[10] Dalkir, K., & Katz, R. (2020). Navigating Fake News, Alternative Facts, and Misinformation in a Post-Truth World. IGI Global.

[11] Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555.

[12] Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., ... & Riedl, M. O. (2022, April). Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-7).

[13] European Commission. (n.d.). The EU Artificial Intelligence Act. Retrieved June, 25, 2023 from https://www.artificial-intelligence-act.com/

[14] Floridi, L. (2023). On good and evil, the mistaken idea that technology is ever neutral, and the importance of the double-charge thesis. *Philosophy & Technology*, *36*(3), 1–5.

[15] Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., Pitliya, R. J., Heins, C., Klein, B., & Millidge, B. (2022). Designing ecosystems of intelligence from first principles. *arXiv Preprint arXiv:2212.01354*.

[16] Hendrycks, D. (2023). Natural selection favors ais over humans. *arXiv preprint arXiv:2303.16200.*

[17] Islam, A. K. M. N., Laato, S., Talukder, S., & Sutinen, E. (2020). Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, *159*. Scopus. https://doi.org/10.1016/j.techfore.2020.120201

[18] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv Preprint arXiv:2307.10169*.

[19] Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1–4), 55–81.

[20] Lance, B. W., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. European Journal of Communication, 33(2), 122–139. Library & Information Science Abstracts (LISA). https://doi.org/10.1177/0267323118760317

[21] Laufer, B., Gilbert, T., Nissenbaum, H., (2023). Optimizations' Neglected Normative Assumptions. ACM FAccT Conference. Retrieved June 25, 2023, from https://www.youtube.com/watch?v=z2stpznzMsc&ab_channel=ACMFAccTConference .

[22] Lewandowsky, S. (2020). The 'post-truth'world, misinformation, and information literacy: A perspective from cognitive science. Informed Societies, 69.

[23] Newfield, C. (2023). How to Make "AI" Intelligent; or, The Question of Epistemic Equality. Critical AI, 1(1–2).

[24] Ramonet, I. (1998). Geopolitics of chaos. Algora Publishing.

[25] Rubin, V. L. (2022). Misinformation and Disinformation: Detecting Fakes with the Eye and AI. Springer Nature.

[26] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.