SPRITE+ Explainer #002

Series Editor: Mark Elliot

Artificial Intelligence and Information Disorder

By John McAlaney, Abir Awad, Luca Viganò, Zeba Khanam, Iain Reid, and Robert Dover

This explainer focuses on artificial intelligence (AI) and information disorder. It first outlines the conceptualisation of information disorder and highlights the impact that AI technologies have had on this phenomenon. The following section then discusses the trust, identity, privacy, and security implications of the intersection between AI and information disorder, before consideration is given to some of the possible future developments in this space.

What is information disorder?

Information disorder refers to information that is shared through society, and which is false, misleading, or harmful [33, 28]. This includes what has been called "fake news"[1] and can be further divided as follows:

- **Misinformation** - the unintentional sharing of false information.
- **Disinformation** - the deliberate sharing of false information, including deepfakes.
- **Malinformation** - information that is factually correct but shared in a harmful way.

This is not a new phenomenon, however recent developments in AI have increased the speed with which information disorder can grow, and the extent to which it can spread. This can happen through several processes: i) aiding in the generation of deepfake items such as images, audio and video [25]; ii) amplifying the dissemination of information disorder items by AI driven recommendation algorithms or AI empowered bot networks [26]; or iii) utilising AI empowered personalisation to target individuals with tailored information disordered items, typically based on behavioural profiling [22]. This behavioural profiling is itself enabled by the information that individuals generate through digital devices, including social media posts and that information may in turn then be used as training data for AI models. It has further been noted that large language models (LLM) demonstrate high levels of sycophancy, in which they are likely to agree with whatever statement a user makes, even when this statement may be factually incorrect [24].

When considering the relationship between AI and information disorder it is important to understand how we process the extensive volume of complex social information that we encounter in our daily lives. As humans we use heuristics to help us do this – these are mental shortcuts that enable us to sift through myriad sources of information to

[1] We prefer the term *false news* since, whilst "fake news" is the popular term, it is contested in the research literature as it does not distinguish between false information that is shared knowingly (disinformation) and that is shared unwittingly (misinformation). Information disorder appears to be becoming the term that is becoming accepted in the literature, although the discourse is developing and still quite fluid.

identify what we believe to be important and truthful [17]. Whilst these heuristics serve a useful evolutionary purpose, they can lead to us to coming to erroneous conclusions, partly through creating cognitive biases: systematic deviations from rationale judgement [16]. These biases can in turn be deliberately exploited by malicious actors. This is something that is already well-documented in the case of social engineering within cybersecurity, in which for example phishing emails use the inclusion of visual cues such as company logos to convince a recipient that a communication is genuine [5].

An example of these heuristics particularly relevant to information disorder is *confirmation bias*, where we are more likely to believe that something is true if it matches our pre-conceptions about the world [27]. An AI empowered system that seeks to create information disorder could target these preconceptions with tailored media content. This is also consistent with established psychological processes such as *false consensus*, in which individuals erroneously believe that those around them agree on an issue [29]. These effects may be further strengthened by the influence of filter bubbles [14] and echo chambers [8]. In the case of filter bubbles, algorithms are used to personalise content to individuals, based on their past clicks and searches. As such they receive selective information that is less likely to challenge their perceptions of the world. Echo chambers on the other hand refers to the tendency of individuals on social media to interact with those who share the same views and beliefs as themselves, resulting in social reinforcement of those views and beliefs.

These conditions enable the spread of information disorder, as well as narrowing the information sources and contrasting viewpoints that an individual may otherwise encounter. It has been suggested that organisations and societies benefit from what is called *cognitive diversity*, in which there are a range of viewpoints and opinions [23]. AI empowered information disorder techniques can however be used to target and amplify specific sub-sets of beliefs to create an illusion that those beliefs are more widespread than they are, potentially invoking societal conflict [6].

Critical issues for TIPSS

Information disorder represents a threat to trust, identity, privacy, security and safety because, fundamentally, it creates uncertainty for individuals about what information is truthful. This can be exploited by adversaries to deceive and manipulate their targets. There are varying conceptualisations of trust, of which one of the most widely used is the idea that it is based upon positive expectations of the intentions or behaviour of another [30]. AI empowered information disorder could undermine this by, for instance, leading people to doubt the intentions behind Government messaging relating to public safety. Identity can also be exploited in several ways, such as through the hijacking of identity to spread information disorder items by use of voice cloning and deepfakes, as well as the use of personal information to tailor misinformation messaging to specific targets. Privacy can be compromised through the ways in which AI is trained on personal data, which can occur without the individual's knowledge or consent [21]. This personal data can then be used for micro-targeting and personalisation of information order items, as discussed previously. In the case of AI empowered malinformation, doxxing at large scales can also occur, where personal details about an individual are combined with false or misleading information, with the intent of causing harm [11]. Security can be threatened using AI empowered information disorder to achieve various goals such as acts of cyberwarfare [15], and through the creation of personalised and naturalistic phishing emails [13]. AI technologies can also be used

to erode trust in legitimate information sources, which can undermine the public's understanding of where and how to seek valid cybersecurity advice [4]. Together, these issues contribute towards concerns around the safety implications of AI on information disorder, with recent research also highlighting the direct impact on mental health issues such as psychosis [9].

AI amplifies the risks of information disorder; however, it also has the potential to mitigate information disorder. This can include using AI to detect and classify misinformation [2; 12]; to moderate false or misleading content on social media platforms and to fact check against trusted databases [1]; and to map how misinformation spreads through networks so that we can better understand these processes [32]. In other words, the same factors that make AI a threat for information disorder – i.e. the speed and scale on which AI can act – can also be used to mitigate and prevent the harms of information disorder. Nevertheless, it must be acknowledged that there are limitations in this approach, such as biases in AI-detection models [31]. There is also a need for ethical oversight in the development of any such models [3].

Future developments

As AI develops it is likely that the abilities it provides to create and amplify information disorder will increase. This could include the emergence of hyper-personalisation, where false news is tailored not just to individual's beliefs but also to their own writing style or preferred tone of voice [36]. There may also come a point where the amount of false news content generated by AI exceeds the amount of factual content created by humans. This could lead to information flooding, where there is more information to be processed than humans can achieve [7], even with the use of heuristics and cognitive biases. The consequences of this are not limited to individuals. It has been noted that for instance that the threats caused by these technologies include undermining democratic processes [16]. However, this information flooding could in turn result in a more comprehensive authentication ecosystems, which can themselves be facilitated through use of AI [17]. In addition, the implementation of proposed changes to regulatory frameworks such as the European Telecommunications Standards Institute technical specification *Securing Artificial Intelligence (SAI): Baseline Cyber Security Requirements for AI Models and Systems* may provide greater clarity and benchmarks around the securing AI system, including those that contribute to information disorder.

Some more specific developments of AI in this space could include the use of AI personas, which are persistent and human-like AI social media accounts that could conduct long-term influence operations in online communities [35]. This could be applied in emerging technologies such as virtual and augmented reality, where individuals could unwittingly directly converse in real-time with an AI agent that has a goal of spreading misinformation [20] with the AI agent performing a personalised and long-term influence operation on the targeted individual.

Overall, there is an arms race over the use of AI to both create and mitigate information disorder [2]. This highlights the need for an ethical framework that ensures a balance between the advancement of AI and the ethical principles and societal needs that relate trust, identity, privacy and security. This includes key considerations such as transparency, accountability, fairness and governance [10]. The implementation of responsible AI practices is essential and should encompass user engagement, training and exchange of knowledge among AI users [34]. Finally, as the technology continues to develop it is important that we explore the human aspect of these developments, so that we can better understand how to identify and mitigate the

threats to trust, identity, privacy, and security, as well as identifying the positive opportunities that these technologies can provide.

References

[1] Augenstein I, Bakker M, Chakraborty T, Corney D, Ferrara E, Gurevych I, Hale S, Hovy EH, Ji H, Larraz I, Menczer F, Nakov P, Papotti P, Sahnan D, Warren G, Zagni G. Community moderation and the new epistemology of fact checking on social media. *arXiv*. 2025;abs/2505.20067.

[2] Bano S, Baig A, Abrejo S. Combating digital misinformation and deepfakes using artificial intelligence: Analyzing the role of AI in detection, content moderation, and public trust in the era of information disorder. *Annu Methodol Arch Res Rev*. 2025 May 5;3(5):78-91.

[3] Bevilacqua M, Berente N, Domin H, Goehring B, Rossi F. The return on investment in AI ethics: A holistic framework. 2024. Available from: https://doi.org/10.24251/HICSS.2024.701.

[4] Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Cebrian M, Schwalbe U, Drexler K. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Oxford: Future of Humanity Institute, University of Oxford; 2018.

[5] Butavicius M, Parsons K, Pattinson M, McCormac A. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. 2016. Available from: https://doi.org/10.48550/arXiv.1606.00887.

[6] Cybenko G, Cybenko AK. AI and fake news. *IEEE Intell Syst*. 2018;33(5):1-5. https://doi.org/10.1109/MIS.2018.2877280.

[7] Danry V, Pataranutaporn P, Epstein Z, Groh M, Maes P. Deceptive AI systems that give explanations are just as convincing as honest AI systems in human-machine decision making. 2022. Available from: https://doi.org/10.48550/arXiv.2210.08960.

[8] Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W. Echo chambers: Emotional contagion and group polarization on Facebook. *Sci Rep*. 2016;6:37825. https://doi.org/10.1038/srep37825.

[9] Dohn'any S, Kurth-Nelson Z, Spens E, Luettgau L, Reid A, Gabriel I, Summerfield C, Shanahan M, Nour MM. Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv*. 2025;abs/2507.19218.

[10] Germani F, Spitale G, Biller-Andorno N. The dual nature of AI in information dissemination: Ethical considerations. *JMIR AI*. 2024;3:e53505. https://doi.org/10.2196/53505.

[11] Goldstein J, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative language models and automated influence operations: Emerging threats and potential mitigations. 2023. Available from: https://doi.org/10.48550/arXiv.2301.04246.

[12] Gondwe G. Can AI outsmart fake news? Detecting misinformation with AI models in real-time. *Emerg Media*. 2025;3(2):252-74. https://doi.org/10.1177/27523543251325902.

[13] Heiding F, Lermen S, Kao A, Schneier B, Vishwanath A. Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. 2024. Available from: https://doi.org/10.48550/arXiv.2412.00586.

[14] Holone H. The filter bubble and its effect on online personal health information. *Croat Med J*. 2016;57(3):298-301. https://doi.org/10.3325/cmj.2016.57.298.

[15] Hunter LY, Albert CD, Rutland J, Topping K, Hennigan C. Artificial intelligence and information warfare in major power states: How the US, China, and Russia are using artificial intelligence in their information warfare and influence operations. *Def Sec*

Anal. 2024;40(2):235-69. https://doi.org/10.1080/14751798.2024.2321736.

[16] Jain S, Spelliscy C, Vance-Law S, Moore S. AI and democracy's digital identity crisis. *arXiv*. 2023;abs/2311.16115.

[17] Kahneman D. Thinking, fast and slow. New York: Farrar, Straus and Giroux; 2011.

[18] Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press; 1982.

[19] Kourtesis P. A comprehensive review of multimodal XR applications, risks, and ethical challenges in the Metaverse. *Multimodal Technol Interact*. 2024;8:98.

[20] Islam MBE, Haseeb M, Batool H, Ahtasham N, Muhammad Z. AI threats to politics, elections, and democracy: A blockchain-based deepfake authenticity verification framework. *Blockchains*. 2024;2(4):458-81. https://doi.org/10.3390/blockchains2040020.

[21] Trinh L, Liu Y. An examination of fairness of AI models for deepfake detection. *arXiv*. 2021. Available from: https://arxiv.org/abs/2105.00558.

[22] Longpre S, Mahari R, Lee AN, Lund C, Oderinwale H, Brannon W, Saxena N, Obeng-Marnu N, South T, Hunter C, et al. Consent in crisis: The rapid decline of the AI data commons. *arXiv*. 2024;abs/2407.14933.

[23] Matz SC, Teeny JD, Vaid SS, Peters H, Harari GM, Cerf M. The potential of generative AI for personalized persuasion at scale. *Sci Rep*. 2024;14(1):4692. https://doi.org/10.1038/s41598-024-53755-0.

[24] Milliken FJ, Martins LL. Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Acad Manage Rev*. 1996;21(2):402-33. https://doi.org/10.5465/amr.1996.9605060217.

[25] Naddaf M. AI chatbots are sycophants — researchers say it's harming science. *Nature*. 2025;647:13-4.

[26] Nasiri S, Hashemzadeh A. The evolution of disinformation from fake news propaganda to AI-driven narratives as deepfake. *J Cyberspace Stud*. 2025;9(1):229-50.

[27] Ng LHX, Carley KM. A global comparison of social media bot and human characteristics. *Sci Rep*. 2025;15(1):10973. https://doi.org/10.1038/s41598-025-96372-1.

[28] Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol*. 1998;2:175-220.

[29] Ricard J, Yañez I, Hora L. A framework for information disorder: Modeling mechanisms and implications based on a systematic literature review. 2025. Available from: https://doi.org/10.48550/arXiv.2504.12537.

[30] Ross L, Greene D, House P. The false consensus effect: An egocentric bias in social perception and attribution processes. *J Exp Soc Psychol*. 1977;13:279-301.

[31] Rousseau DM, Sitkin SB, Burt RS, Camerer C. Not so different after all: A cross-discipline view of trust. *Acad Manage Rev*. 1998;23(3):393-404.

[32] Trivedi A, Suhm A, Mahankal P, Mukuntharaj S, Parab MD, Mohan M, Berger M, Sethumadhavan A, Jaiman A, Dodhia R. Defending democracy: Using deep learning to identify and prevent misinformation. *arXiv*. 2021;abs/2106.02607.

[33] Wardle C, Derakhshan H. Information disorder: Toward an interdisciplinary framework for research and policy making. 2017. Available from: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

[34] Williamson SM, Prybutok V. The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*. 2024 Jun;15(6):299.

[35] Yang KC, Singh D, Menczer F. Characteristics and prevalence of fake social media profiles with AI-generated faces. *arXiv*. 2024;abs/2401.02627.

[36] Zugecova A, Macko D, Srba I, Móro R, Kopal J, Marcincinova K, Mesarcík M. Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. *arXiv*. 2024;abs/2412.13666.