Generative AI as a Disinformation Tool: "The Hidden Sound of Things Approaching"

https://doi.org/10.22151/politikon.61.CON3

Marios D. DIKAIAKOS
University of Cyprus
mdd@ucy.ac.cy

Abstract

The rapid spread of Large Language Models (LLMs) is transforming how individuals access and interpret information, increasing societal exposure to sophisticated forms of misinformation. As generative AI becomes a central information gatekeeper, it expands the attack surface for Foreign Information Manipulation and Interference (FIMI), enabling scalable data contamination, alignment manipulation, and realistic synthetic media. These dynamics weaken critical evaluation skills and amplify latent model biases, paralleling the long-term cognitive effects of traditional disinformation. Using lessons from science misinformation, the paper argues for robust regulation, transparent training practices, and integrated resilience strategies to mitigate emerging systemic risks posed by generative AI.

Keywords: Artificial Intelligence and Democracy; Generative AI; Large Language Models; Misinformation; Digital Resilience; Foreign Information Manipulation and Interference (FIMI)

This essay argues that the exponential adoption of Large Language Models (LLMs) and Artificial Intelligence (AI) chatbots is transforming these systems into a dominant mediator between individuals and their digital information environment. Echoing C. P. Cavafy's verse that "the hidden sound of things approaching" is heard only by the "wise" who "listen reverently," while the many "hear nothing whatsoever" (Cavafy 1992), the hype around generative AI's onslaught obscures the coalescence of its risks. This obliviousness renders democratic societies increasingly vulnerable to LLM-driven misinformation, a threat that becomes particularly acute in light of growing concerns over the potential weaponization of LLMs in Foreign Information Manipulation and Interference (FIMI) operations. Such campaigns, orchestrated by foreign actors, aim at eroding public trust, distorting civic discourse, and weakening democratic governance for strategic political gain (Elsner, Atkinson, and Zahidi 2025; UK Government 2025; Hassoun et al. 2024; Bergmanis-Korāts et al. 2024).

Addressing misinformation within and through LLMs is anything but straightforward. Although several leading models incorporate so-called "guardrails," i.e., mechanisms intended to align LLM outputs with established societal norms and safety constraints, the boundaries of what constitutes "acceptable content" remain deeply contested and vary significantly across cultures, communities, and social strata. Moreover, although LLMs generate compelling content that appears factual, this content may at times be entirely fabricated or subtly shaped by biases embedded in LLM training data. Also, recent studies demonstrate that guardrails can be bypassed, and LLMs can be fine-tuned to deliberately shift their alignment toward specific objectives (Hsiung et al 2025; Paschalides, Pallis, Dikaiakos 2025). These developments, combined with the impressive capacity of AI diffusion models to fabricate realistic images, audio, and video ("deepfakes"), create further opportunities for malicious actors who wish to exploit such technologies and significantly expand the scale, speed and situational awareness of known disinformation Tactics, Techniques, and Procedures (TTP), adapting them effectively to various cultural, political or situational contexts (Bergmanis-Korāts et al. 2024). Therefore, malicious actors who already exploit the business models and algorithmic dynamics of computational advertising across digital platforms, social media, and messaging apps (Bergmanis-Korāts and Haiduchyk 2024), can now leverage powerful generative AI tools to amplify their TTPs with unprecedented precision and persuasiveness, micro-targeting content toward specific groups and objectives continuously and at a minimal cost.

Looking beyond these alarming scenarios, the rapid adoption of powerful LLM-based tools and their integration in human activities across diverse domains, from education and scientific research to journalism and policymaking, will greatly expand the attack surface of FIMI campaigns. As LLM-based systems become common information gatekeepers, they reduce our direct access to original sources and discourage active and critical evaluation. This may undermine the ability of humans, and especially younger generations, to critically appraise information, hence becoming more vulnerable to the hidden biases and inaccuracies latent in LLM outputs (The Economist 2025; Lowe 2025). Such biases can arise from: (a) intentional or inadvertent data contamination in datasets used for LLM training, distorting content quality and undermining reliability; (b) covert alignment, where LLM outputs are subtly shaped by the business model(s) and strategic objectives of their providers while maintaining the appearance of neutrality; and (c) deceptive alignment, whereby models comply with user expectations while secretly prioritizing undisclosed goals and influencing perceptions in subtle yet consequential ways (Carranza et al. 2023).

In this context, adversarial actors, mirroring the dynamics of political propaganda or manipulative public relations, could exploit LLM vulnerabilities to interfere with the LLM training process. By systematically contaminating training datasets with false or biased examples, adversarial inputs could affect an LLM's classification mechanism, reshaping the criteria it uses to interpret and respond to queries. Similar to the ultimate objective of traditional disinformation campaigns, which go beyond the dissemination of falsified facts and aim at changing their targets' opinion formation and long-term behavior (Rid 2021; Bola and Papadaki 2021), the goal here would be to subvert the LLM's foundational decision-making framework, ensuring skewed judgments even in scenarios where input data is otherwise accurate. For instance, adversarial training data could subtly prime LLMs to prioritize certain ideological narratives or commercial interests, amplifying biases or inaccuracies through latent algorithmic preferences. Conversely, structured and transparent training with rigorously curated datasets can act as a learning-enhancing force, reinforcing robust alignment with transparency and accuracy while mitigating the risks of adversarial interference.

Taking as an example science disinformation (NASEM 2024), we could imagine malicious actors who seek to undermine particular scientific theories, engaging in targeted campaigns to pollute the information sources used to train LLM systems by creating fake scientists' profiles, publishing bogus scientific reports in open scientific archives, posting fictitious datasets in data repositories, creating inauthentic citation cartels (Pérez-Neri, Pineda, and Sandoval 2022) to establish fake credibility in bogus articles (Lockwood 2020), etc. Some of these fraudulent practices are already exploited by fraudsters and predatory journals seeking financial or reputational gain. However, such cases typically operate on a limited scale and can still be detected and countered by scientific experts through peer review, retractions, and institutional oversight mechanisms. The concern is that LLMs could automate and massively scale these deceptive behaviors, producing plausible but false research outputs that overwhelm traditional safeguards. Automatically identifying such practices within the vast, heterogeneous datasets used to train LLMs remains fraught with uncertainty, let alone reliably filtering them to prevent harmful outputs. The risks of failure are severe: unchecked biases or disinformation could propagate at scale, with damage persisting long after initial exposure.

A cautionary example is the fraudulent 1998 *Lancet* study that falsely linked the MMR vaccine to autism (Deer 2011). Just as this case illustrates, once flawed or malicious information enters public discourse, it can persist despite formal correction (Rao and Andrade 2011), producing long-lasting and damaging effects. Similarly, contaminated or biased training data in LLMs can shape model outputs in unpredictable and potentially harmful ways. These parallels underscore the pressing need for safeguards that go beyond surface-level detection to confront the latent diffusion of harmful ideas within systems trained on imperfect data. Strengthening protections to reduce the susceptibility of LLM training to "polluted" datasets and enhancing the resilience of LLM outputs against manipulation remain challenging problems.

These challenges would be tackled more aggressively by the industry if AI regulatory frameworks prioritized responsible development and enforcement of AI technologies and applications (Judge, Nitzberg, and Russell 2024). However, major economies like the US and EU appear recently to frame regulation as an impediment to innovation rather than a necessity (Roose 2025), and this approach risks entrenching systemic vulnerabilities in LLMs, allowing flawed or malicious content to persist in training pipelines and propagate through downstream applications. Without strong guidelines that reconcile agility with accountability, the unchecked escalation of Generative AI adoption risks could surpass the societal harms exemplified by decades of vaccine misinformation emanating from the fraudulent 1998 *Lancet* study. But even strong regulatory guidelines cannot be sufficient. European countries must urgently design and continuously reinforce comprehensive, adaptive, and innovative strategies that will integrate robust regulation of LLM services with deep educational reforms and sustained public awareness initiatives, ensuring that societal resilience evolves in step with the accelerating pace of AI innovation.

References

Bergmanis-Korāts, G., and T. Haiduchyk. 2024. *Social Media Manipulation for Sale: 2024 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement*. NATO Strategic Communications Centre of Excellence.

Bergmanis-Korāts, G., G. Bertolin, A. Pužule, and Y. Zeng. 2024. *AI in Support of StratCom Capabilities*. NATO Strategic Communications Centre of Excellence.

Bola, C., and K. Papadaki. 2021. "Digital Propaganda, Counter Publics, and the Disruption of the Public Sphere: The Finnish Approach to Building Digital Resilience." In *The World Information War: Western Resilience, Campaigning, and Cognitive Effects*, edited by T. Clack and R. Johnson. Routledge.

Carranza, A., et al. 2023. "Deceptive Alignment Monitoring." *arXiv* 2307.10569.

Cavafy, C. P. Collected Poems. Revised ed. Translated by Edmund Keeley and Philip Sherrard. Edited by George Savidis. Princeton, NJ: Princeton University Press, 1992.

Deer, Brian. 2011. "How the Case Against the MMR Vaccine Was Fixed." *BMJ* 342:c5347. https://doi.org/10.1136/bmj.c5347.

Elsner, M., G. Atkinson, and S. Zahidi. 2025. *The Global Risks Report 2025, 20th ed.* World Economic Forum.

Hassoun, A., A. Abonizio, K. Osborn, C. Wu, and B. Goldberg. 2024. "The Influencer Next Door: How Misinformation Creators Use GenAI." *arXiv* 2405.13554.

Hsiung, L., Pang, T., Tang, Y-C., Song, L., Ho, T-Y., Chen, P-Y., Yang, Y. "Why LLM Safety Guardrails Collapse After Fine-tuning: A Similarity Analysis Between Alignment and Fine-tuning Datasets." *arXiv* 2506.05346v1.

Judge, B., M. Nitzberg, and S. Russell. 2024. "When Code Isn't Law: Rethinking Regulation for Artificial Intelligence." *Policy and Society*. https://doi.org/10.1093/polsoc/puae020.

Lockwood, M. 2020. "Editorial: Citation Malpractice." *Proceedings of the Royal Society A* 476: 20200746.

Lowe, D. 2025. "An Evaluation of 'Deep Research' Performance." *In the Pipeline* (blog). https://www.science.org/content/blog-post/evaluation-deep-research-performance.

NASEM, National Academies of Sciences, Engineering, and Medicine. 2024. *Understanding and Addressing Misinformation About Science*. Washington, DC: National Academies Press.

Paschalides, D., Pallis, G., Dikaiakos, M.D. 2025. "Adopting Beliefs or Superficial Mimicry? Investigating Nuanced Ideological Manipulation of LLMs." *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media* (ICWSM 2025).

Pérez-Neri, I., C. Pineda, and H. Sandoval. 2022. "Threats to Scholarly Research Integrity Arising from Paper Mills: A Rapid Scoping Review." *Clinical Rheumatology* 41 (7): 2241–48. https://doi.org/10.1007/s10067-022-06198-9.

Rao, T. S., and C. Andrade. 2011. "The MMR Vaccine and Autism: Sensation, Refutation, Retraction, and Fraud." *Indian Journal of Psychiatry* 53 (2): 95–96. https://doi.org/10.4103/0019-5545.82529.

Rid, Thomas. 2021. *Active Measures*. Picador USA.

Roose, Kevin. 2025. "5 Notes From the Big A.I. Summit in Paris." *New York Times*, February 10.

The Economist. 2025. "The Danger of Relying on OpenAI's Deep Research." February 13.

UK Government. 2025. "The Bletchley Declaration by Countries Attending the AI Safety Summit."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.