The Role of AI Agents for Online Information Disorders

The 2025 World Economic Forum has identified AI-generated misinformation and disinformation as the most urgent short-term global risks. LLM-powered agents can exacerbate these issues, amplifying the spread of convincing false narratives and undermining public trust.

By Francesco Pierri

DOI: 10.1145/3729511

OPEN ACCESS

Since the 2016 U.S. Presidential election, interest in the dynamics of misinformation—false or misleading information shared without harmful intent—and disinformation—intentionally fabricated or manipulated information, often spread through coordinated campaigns of real and fake accounts and bots—has surged in academic and public discourse [1]. More recently, the rise of large language models (LLMs) has introduced new threats by enabling the creation of sophisticated, scalable, and highly persuasive content.

Early research in the "science of fake news" has shown that false content tends to be disproportionately shared by a small subset of users—often older individuals with conservative political views. Echo chambers—insular online communities where individuals are primarily exposed to information that aligns with their preexisting beliefs—can facilitate the spread of misinformation by reinforcing and validating false narratives within these groups. These dynamics create a feedback loop, where false narratives gain credibility through repeated exposure.

Social bots further exacerbate this problem by automating the dissemination of misinformation on a massive scale. These bots can artificially boost the visibility and perceived legitimacy of false information, manipulate trending topics, and engage with users to perpetuate misleading content.

The implications of these dynamics are particularly alarming during political elections and geopolitical events, where access to reliable information is critical for informed decision-making. In such contexts, the rapid spread of misinformation and disinformation can manipulate public opinion, polarize voters, and undermine trust in democratic institutions. Geopolitical

Image by Snopek/Shutterstock.com

crises, such as conflicts or international disputes, are also vulnerable to information manipulation aimed at shaping global narratives or destabilizing opponents. The integration of LLMs into this ecosystem heightens these risks by enabling the rapid production of tailored, contextually convincing narratives, making it increasingly difficult to differentiate between authentic and fabricated content during pivotal moments.

THE IMPACT OF MISINFORMATION DURING GLOBAL CRISES

The recent COVID-19 pandemic and the 2022 Russian invasion of Ukraine are notable examples of how misinformation and disinformation can have profound consequences on the real world. During the pandemic, the spread of unreliable information in digital environments fueled "infodemics," overwhelming public discourse with falsehoods that undermined trust in health measures and institutions. Similarly, the invasion of Ukraine saw an explosion of coordinated disinformation campaigns aimed at manipulating narratives, sowing division, and influencing global perceptions of the conflict.

At the Indiana University Observatory on Social Media, we analyzed the first year of the COVID-19 infodemic on two major social media platforms, Twitter (now X) and Facebook, finding significant differences in the prevalence of popular low-credibility sources and suspicious videos between the two platforms [2]. On both platforms, however, a small number of accounts and pages wielded disproportionate influence as "superspreaders" of misinformation, paradoxically verified by the platforms themselves. We observed similar patterns as we monitored and visualized the spread of vaccine misinformation on Twitter in the United States through our public dashboard CoVaxxy [3].

In particular, we found that while low-credibility content was less prevalent than mainstream news, it often achieved similar or greater reshare volumes compared to trusted sources like the Centers for Disease Control and Prevention CDC and WHO. Alarmingly, around 800 verified "superspreaders" were responsible for 35% of all misinformation reshares daily. These results highlighted the persistent influence of a small group of repeat misinformation spreaders, likely motivat-

ed by financial incentives, even amid deplatforming efforts.

The prevalence of online misinformation during the pandemic was particularly worrisome, as exposure to false and misleading content had been linked to heightened health risks and a surge in vaccine hesitancy, which is a wide concept ranging from delays in vaccination to outright refusal to get vaccinated. Geolocating users sharing vaccine misinformation tweets and leveraging daily surveys administered by Facebook to measure individuals' intentions to get vaccinated in the United States, we found that online misinformation was negatively associated with vaccination uptake rates (and positively with vaccine hesitancy), in line with previous research that tested this link in a controlled experiment. These associations remained significant even after controlling for political affiliation, demographics, and socioeconomic factors. While vaccine hesitancy was closely linked to Republican vote share, the influence of online misinformation on hesitancy was most pronounced in Democratic counties. This research [4], where we advocated for interventions to support better-informed health decisions, was later referenced in the 2023 "Economic Report of the President."

During my research stay at the University of Southern California, we conducted a longitudinal analysis of misinformation and propaganda on Facebook and Twitter in the early months of the 2022 Russian invasion of Ukraine [5]. Using a dataset comprising nearly 20 million Facebook posts and more than 250 million tweets, we examined the prevalence, spread, and moderation of Russian propaganda and low-credibility content compared to high-credibility sources. Our findings revealed that while Russian propaganda declined after the invasion due to platform policies, European sanctions, and Russia's ban on certain platforms, low-credibility content remained steady and continued to generate significant engagement. During the same period, we observed spikes in account creation which are typical of global events such as conflicts and political elections [6], with many of these new accounts being suspended shortly after for violating Twitter's policies as they exhibited distinct behavioral features, including excessive use of replies, higher activity levels, and frequent dissemination of harmful or spam content. We also highlighted once again the disproportionate role of a small number of "superspreaders," many of whom were verified accounts, in amplifying misinformation. These accounts were responsible for the majority of interactions and retweets related to unreliable content. Despite efforts by platforms to moderate misinformation, only 8–15% of posts and tweets linking to unreliable sources were removed.

RISKS OF NOVEL GENERATIVE AI

Ever since the launch of ChatGPT, the rise of LLM-powered agents (hereby "AI agents") has raised concerns about their potential to exacerbate issues related to misinformation, disinformation, and influence operations [7]. Bad actors can exploit AI agents to scale malicious activities such as spreading misinformation, generating harmful content, or orchestrating manipulative campaigns by exploiting algorithmic and socio-cognitive vulnerabilities. Likewise, everyday users may inadvertently use AI agents in ways that propagate false claims or produce biased and offensive outputs, intensifying risks without intentional malice.

LLMs often struggle with factuality, generating content that is confidently incorrect or misleading. They may produce outputs without relying on credible or verifiable sources, making it difficult to discern truth from falsehood. Their polished and confident delivery can make false information appear credible, while their widespread availability enables even users with minimal expertise to generate large volumes of harmful or misleading content. Additionally, the public's perception of AI as neutral and authoritative can lend undue credibility to AI-generated outputs, even when incorrect, and current methods for evaluating their factuality and reliability remain insufficient, particularly in complex or evolving contexts.

Beyond factuality, LLMs also pose significant security and manipulation risks. They can be used to craft highly personalized phishing messages, ha-

rassment, or targeted misinformation, while their ability to mimic the tone of trusted individuals or institutions makes impersonation easier. Adversaries can fine-tune or prompt LLMs to bypass content detection systems, allowing harmful content to proliferate undetected. Moreover, LLMs can generate realistic fake profiles at scale, which may amplify false narratives, manipulate public opinion, or facilitate coordinated influence operations.

The European Union has expressed particular concern about the risks posed by AI-driven misinformation and disinformation, leading to the introduction of regulatory frameworks such as the AI Act and the Digital Services Act (DSA). These measures mandate stringent accountability, transparency, and safety standards for both platforms and AI providers, aiming to mitigate the potential misuse of AI technologies. The AI Act focuses on regulating high-risk AI systems, including those capable of generating or amplifying harmful content, by requiring robust risk assessments, clear labeling, and strict compliance with ethical guidelines. Meanwhile, the DSA emphasizes platform responsibility, compelling digital platforms to monitor, report, and mitigate the spread of disinformation, particularly through algorithmic transparency and content moderation practices. Together, these initiatives reflect the EU's proactive stance in addressing the evolving challenges of AI-driven online harms while balancing innovation with societal safety.

LLM-POWERED AGENTS TO SIMULATE ONLINE HUMAN BEHAVIOR

Despite the risks posed by LLMs, generative AI agents offer unprecedented opportunities for simulating user behavior in digital environments [8]. Their ability to autonomously operate, dynamically adapt to complex scenarios with human-like planning, and seamlessly interact with both synthetic agents and real users positions them as uniquely suited for modeling online interactions. AI agents can be personalized to impersonate diverse user demographics, capturing a wide range of behaviors, preferences, and interaction styles, thereby enhancing the relevance and fidelity of simulations in studying digital ecosystems.

The rapid spread of misinformation and disinformation can manipulate public opinion, polarize voters, and undermine trust in democratic institutions

Still, research on LLMs interacting with each other remains in its infancy, relying on simplistic approaches to network construction that fail to capture the complexity and evolution of real-world social networks. Simulating social media ecosystems with LLMs faces significant technical challenges, which impact the accuracy and reliability of simulations. Consistency is a major concern, as LLMs often struggle to maintain coherent behavior across extended interactions due to limited memory, leading to fragmented or shallow simulations. Unpredictable behavior further complicates modeling, particularly in scenarios requiring alignment with specific user archetypes. Additionally, LLMs are vulnerable to jailbreaking and adversarial attacks, where malicious inputs or manipulative prompts exploit model weaknesses to bypass safeguards or induce unintended behaviors.

Indeed, AI agents can have a dual-edged impact on online information disorders. They can amplify false narratives by generating highly credible, misleading content that mirrors human-created material and is difficult to detect. Conversely, they show promise in mitigating misinformation by generating counter-narratives or serving as tools to correct false beliefs. However, reliable methods for systematically evaluating the societal impact of AI-generated misinformation and disinformation campaigns are lacking, making it difficult to measure their influence. Likewise, research on leveraging LLMs for interventions is still in its early stages, with much unknown about their long-term effectiveness and ethical implications.

CONCLUSION

The integration of LLM-powered agents into the digital ecosystem marks a turning point in both the propagation and mitigation of misinformation and disinformation. While these technologies present significant risks by enabling the creation and dissemination of false narratives at an unprecedented scale, they also offer potential solutions such as counter-narratives and enhanced simulations for studying online behavior. To navigate this dual-edged impact, robust regulatory frameworks, interdisciplinary research, and global collaboration are essential. By advancing methods to evaluate and address AI-driven harms while leveraging their capabilities for positive applications, we can work toward a safer, more informed, and ethically sound digital future.

References

[1] Lazer, D. M. et al. The science of fake news. *Science*. 359, 6380 (2018), 1094–1096.

[2] Yang, K. C. et al. The COVID-19 infodemic: Twitter versus Facebook. *Big Data & Society* 8, 1 (2021).

[3] Pierri, F. et al. One year of COVID-19 vaccine misinformation on Twitter: Longitudinal study. *Journal of Medical Internet Research* 25 (2023), e42227.

[4] Pierri, F. et al. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*. 12, 5966 (2022).

[5] Pierri, F. et al. Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM, New York, 2023, 65–74.

[6] Pierri, F. et al. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Science* 12, 43 (2023).

[7] Augenstein, I. et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* 6, 8 (2024), 852–863.

[8] Park, J. S. et al. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, 2023, 1–22.

Biography

Francesco Pierri is an assistant professor in the Data Science group within the Department of Electronics, Information and Bioengineering (DEIB) at Politecnico di Milano, and affiliated with Indiana University Bloomington's Observatory on Social Media (OSOME). His research combines computational social science and AI, using data-driven methods to study large-scale online phenomena, with a focus on the transformative impact of generative AI on digital information ecosystems.

1528-4972/25/03 $15.00