Promoting Peace and Security

Horizon Insights

Quarterly Journal by Beyond the Horizon ISSG – Volume 7 Issue 1 - 2024 (Jan - Mar)

AI Enhanced Disinformation: State of Play

Markus Neuvonen

To cite this article:

Neuvonen, M. (2024). AI enhanced disinformation: State of play. Horizon Insights, 7(1), 12–15. https://doi.org/10.31175/hi.2024.01.02

Link to this issue of the journal: https://doi.org/10.31175/hi.2024.01

Horizon Insights Journal Homepage: https://behorizon.org/horizon-insights/

ISSN: 2593-3582 (printed), 2593-3590 (online)

Submit your article

AI Enhanced Disinformation: State of Play

Markus Neuvonen*

Introduction

This article builds on insights from an expert event held in October 2024 in Helsinki, organized by Faktabaari (FI) as part of the EU-CERV funded project Immune 2 Infodemic 2. The event brough together participants from governments, media, security agencies and academia to establish a shared understanding of AI-enhanced disinformation and to promote cross-sector collaboration in mitigating societal harms.

The article outlines the primary risks and challenges posed by AI-enhanced disinformation, highlighting the urgency of coordinated responses and regulatory measures across sectors and governance levels. With the rapid development of AI technologies, these systems have become increasingly embedded in the information ecosystem, they accelerate the production and spread of disinformation, undermining democratic resilience, public trust, and shared epistemic foundations. Six key areas are explored in the article: declining trust in institutions, hyper-targeted disinformation, unaccountable tech governance, media literacy fragility, regulatory gaps, and cognitive security.

The current situation of AI and disinformation

This topic is approached from three different viewpoints—security, media, and democracy. The analysis explores six overarching thematic areas:

1. Degradation of Trust and Information Reliability: AI technologies facilitate low-cost, large-scale dissemination of disinformation, exacerbating societal polarization and undermining public trust in institutions. This trend impacts democratic processes, consumer trust, and corporate decision-making, with rising concerns about deliberate market manipulations.

2. Emerging Risks of AI-Driven Disinformation: AI enables hyper-localized, microtargeted disinformation campaigns, deepfake usage in fraud and political manipulation, and the creation of AI "experts" and influencers. Large Language Models (LLMs) face significant vulnerabilities, particularly regarding the potential corruption of their training data.

3. Digital Power and Corporate Ethics Tech: Corporations' opaque practices and global reach limit governmental oversight. Their role in amplifying microtargeting during elections and collaborating with undemocratic regimes poses threats to democratic values and data privacy.

4. Media Literacy Education and Public Resilience Strengthening: Media literacy and MIL education across all demographics is critical for combating the effects of disinformation. Education systems must prioritize critical thinking and information evaluation skills. Countries such as Norway and Estonia offer best practice examples of integrated approaches to digital literacy. The Faktabaari Digital Information Literacy (DIL) model also serves as a promising framework for building societal resilience.

5. AI Regulation and Policy Participants: Proactive, human-centric regulatory frameworks are urgently needed, particularly on the EU level. Key concerns include the misuse of AI for illicit purposes and the absence of safeguards to prevent violations of ethical norms. Policy efforts must anticipate technological developments and embed accountability mechanisms from the outset.

6. Cognitive Security: AI tools risk distorting shared realities by contributing to information overload, weakening critical thinking, and enabling targeted psychological manipulation. These effects pose long-term risks to democratic governance and individual cognitive autonomy, demanding both policy and societal responses to safeguard mental and civic integrity.

1. Degradation of trust and the crisis of reliability

AI amplifies and accelerates existing challenges within democracies. AI tools can be—and have been—used to construct false narratives, aiming to erode democratic values and undermine public trust in key institutions such as the media, judiciary, political systems, and education. They are a cost-effective way to create and spread these narratives in multiple media formats, which makes them more sustainable and resilient to refutation.

Some of the challenges facing today's media landscape may stem from the underlying business models of media corporations, which prioritize speed and competition—often at the expense of reliability and contextual depth. As the business model relies on freelance journalists, concerns have been raised about varying levels of professionalism, ethical standards, and long-term commitment. In some cases, financial pressures have led freelancers to adopt influencer-like roles, prioritising visibility over journalistic integrity.

These reliability issues can for example be observed in Russian propaganda narratives entering the mainstream

* Senior MIL/DIL specialist at FaktaBaari

media or manipulated headlines. This is crucial since both local news media and national public broadcasters play a central role in fostering public trust and ensuring access to reliable information.

Challenges to information reliability carry significant economic implications, particularly as AI-enabled scams and fraudulent activities become increasingly prevalent. Several high-profile cases illustrate this trend. In one ongoing case that began in 2020, synthetic AI-generated profiles featuring fabricated faces were used in social engineering schemes to extract confidential information and promote false investment opportunities. This has resulted in financial losses and diminished trust in professional networking platforms, demonstrating how AI-generated content can undermine consumer confidence. Another notable case occurred in 2019, when a UK-based energy firm was defrauded after scammers used AI-generated voice cloning to impersonate the German CEO of the firm's parent company. A British executive was deceived into transferring 220,000 to a Hungarian bank account. The deep-fake voice was realistic enough to bypass normal verification, demonstrating how AI can be used to commit cross-border financial fraud with direct economic consequences. Overall, corporate decision-making relies heavily on reliable and verifiable information, yet current means of verifying AI-generated data remain limited. Moreover, AI tools can be exploited for purposes such as deliberate market manipulation.

When it comes to degradation of trust and the crisis of reliability, a key challenge for democratic societies is to safeguard democratic values and institutional trust in environments where AI are deliberately deployed to polarize public discourse and erode confidence in credible sources of information. In response, the notion of a "seedbank of factual knowledge" has emerged–a conceptual safeguard aimed at preserving reliable information and protecting the integrity of the information eco system.

2. Specific risks and challenges caused by disinformation and manipulation

AI systems are associated with several specific risks related to disinformation and manipulation. One of these is that it enables creation and spreading "hyper localised" disinformation; disinformation that is microtargeted to extremely local information bubbles, such as specific neighbourhoods, hobby groups and companies. With this strategy, combining AI powered tools and widely tailored disinformation campaigns, it causes mayhem and influences political elections in a cost-effective manner, often faster than governments are able to respond.

Another emerging issue is the Large Language Models (LLMs), that raise particular concerns. These models are vulnerable to manipulation, distortion, and both accidental and deliberate bias. Due to the opacity of these models and the frequent lack of transparency regarding training data, the widespread use of LLMs present significant threats to the reliability of information and the integrity of democratic discourse. One notable risk involves the deliberate corruption of training dataset and archival sources-potentially orchestrated by state actors or groups seeking to strategically distort foundational knowledge. Particular attention should be given to influence effort originating from Russia and China, whose approaches are often rooted in value-centric manipulation aimed at reshaping global narratives over time.

The rapid advancement of AI technologies is paving the way for the development of autonomous and dynamically adaptive disinformation systems. These involve AI-driven bots that react to real-time events and adjust disinformation narratives accordingly. With the help of generative tools, these systems can simultaneously create persuasive content across multiple media formats and channels. They are also capable of fabricating seemingly credible websites within seconds to amplify and legitimise false messages.

Advancements in AI have also enabled the rise of deepfakes and synthetic video technologies. AI-generated influencers and "experts" are increasingly used to shape public opinion, raising serious concerns about the credibility of expert voices in social media content. Similarly, fake product demonstration videos are becoming more common, further undermining consumer trust and the reliability of visual information online.

The emergence of AI tools has introduced the prospect of automated disinformation production, raising significant concerns given that countermeasures often lag behind due to inertia and slowness. AI-based fact-checking tools represents promising development in efforts to counter disinformation, but with limitations. These systems share many of the same vulnerabilities as other AI-based tools, including exposure to both inadvertent errors and intentional manipulation. Risks such as data contamination, algorithmic bias, and limited transparency continue to challenge their reliability.

3. Digital power and the role and ethics of big tech corporations

Major technology companies have accumulated digital power through their control of social media platforms, search engines, cloud services and AI tools. These corporations control the infrastructure through which information circulates, allowing them to influence public opinion and decision-making to a degree that few democratic institutions or traditional media outlets can rival.

A key concern is the disproportionate political influence these companies hold, paired with their often lack of transparency and limited accountability. By creating dependencies on AI tools these corporations have unprecedented influence on societies on all levels, as after the introduction of efficient AI solutions opting out of them and falling back becomes increasingly difficult if not impossible.

Another key risk associated with these corporations is their apparent willingness to comply with, or even support, undemocratic regimes and agendas—often framed in terms of protecting "freedom of expression". Algorithmic influence on democratic elections has already been witnessed across the globe, as well as using the amassed user data for the purposes of microtargeting under elections.

The existence of this massive data collection by social media giants alone poses a risk despite increasingly strict demands on data privacy. The emergence of AI may eventually enable users to indirectly bypass these data privacy regulations by data combination, making for example the building of "digital twins" more and more usable for both legitimate and clandestine or malevolent purposes.

4. Media and digital information literacy and AI-powered disinformation

In the face of the growing information manipulation and disinformation, media and digital literacy serve as society's most critical safeguard. As such, there is a general concern about the state of media literacy across all age groups. Critical thinking and information evaluation skills are increasingly regarded as essential civic competences, and their inclusion in education—from early years through adulthood—should be considered a societal priority. A narrow focus on technical digital skills alone has proven inadequate in preparing individuals to navigate complex and manipulative information environments.

Formal education plays a central role in developing the analytical and evaluative skills necessary for navigating the age of AI. However, older segments of the population-particularly adults and the elderly are at heightened risk of being left behind in this regard. These groups are typically beyond the reach of formal education systems and must rely instead of self-directed learning, mass media and civil society initiatives. On the national level, Norwegian and Estonian has offered promising approaches on media literacy education, particularly through their use of gamified and digital approaches.

At the same time, the integration of AI in educational settings raises serious concerns. One of the most pressing is the potential erosion of critical thinking, especially among younger students. Excessive reliance on AI tools may weaken cognitive development and reduce the motivation to engage in independent thought and problem-solving. This highlights the urgent need to cultivate "AI literacy"—a set of skills that empowers students to use artificially and responsibly, rather than passively depending on it. In addition, AI technologies are increasingly being misused in harmful ways, including grooming, online sexual exploitation, and cyber bullying, which further underscores the need for protective and educational measures. Materials such as Faktabaari's DIL and AI Literacy Guides can be an example of necessary initiatives for promoting these skills and supporting teachers in promoting them.

Overall, the public's vulnerability to manipulate content poses a serious threat to democracy. Addressing the growing spread of unreliable, misleading and polarizing information requires a unified, strategic approach that bring together actors from the public, private and civil society sectors. However, the absence of unified national strategies, delays in decision-making, limited cross-sectoral and international collaboration, and a prevailing silo mentality within public institutions represent significant obstacles—and consequently major risks—in effectively countering disinformation and its broader societal impacts. There is a need for a shift from reactive to proactive measures and increased collaboration across the field.

5. AI regulation and international issues

The regulation of AI continues to be predominantly reactive, highlighting the urgent need for more strategic, anticipatory, and proactive approaches—both within and beyond the EU. Given that the development and deployment of AI are concentrated in the hands of a few large multinational corporations, national-level regulation alone is insufficient. Similar challenges exist in the media landscape, where national broadcasters have limited capacity to influence dominant global platforms such as Meta and Google.

In this context, the development of EU-level regulation and a coordinated regulatory network becomes not only necessary but essential. Equally important is the advancement of robust technological standards and regulatory frameworks that can guide the responsible use of AI, particularly within the private sector, where many of the most powerful tools are developed and deployed.

Additionally, safeguarding human-centric values in AI development is of critical importance, particularly given the ethical risks involved. There are clear incentives to weaken or bypass ethical safeguards in AI systems, and the practice of "jailbreaking" AI. This includes removing built-in restrictions to enable clandestine or unethical use and is already widespread among certain user groups. This includes use in contexts such as harassment, criminal activity, terrorism, and the distribution of child pornography. These boundaries must be safeguarded more thoroughly and forced through stricter regulation. Legal implications for jailbreaking AI:s should be considered at the very least at national level.

6. Cognitive security

The term "cognitive security" has increasingly been used to describe the long-term impact of disinformation on individuals' worldviews and values. It encompasses the conditions under which individuals retain a coherent and

autonomous understanding of reality. As AI is increasingly used to produce and tailor disinformation, the possibility of a shared epistemic ground—on which democratic processes depend—faces destabilization.

The strategic use of AI to overfeed the social media with nonsensical or overwhelming content can lead to information fatigue, passivity, and emotional reactivity. As a result, individuals' ability to critically assess information becomes overloaded and diminished, which then result in a heightened susceptibility to manipulation. Similarly to the Russian Operation Overload in 2023-24, which aimed to overcrowd Western fact-checkers with AI-generated disinformation and related fact-check requests, this form of deliberate information pollution can be weaponized on a society-wide scale. At the same time, the overwhelming volume of content may push individuals and institutions to rely more heavily on AI tools.

In this context, citizen's ability to critically evaluate information and identify trustworthy sources remain the final safeguard for preserving intellectual integrity and autonomy of the citizenry. Cognitive security is increasingly threatened by AI-driven psychological manipulation, where vast user data sets are used to influence individuals' sense of identity, values, and perception of reality. Through the construction of so-called "digital twins"- detailed behavioural models based on personal and identity-linked data- AI systems can simulate individual responses with high precision. This enables the emergence of reactive disinformation: content dynamically tailored for exploit emotional triggers and cognitive vulnerabilities in real time.

Such developments significantly expand the scope of political and ideological influence. AI tools have already demonstrated the ability to reinforce conspiracy beliefs and shift worldviews, even among seemingly resistant individuals. As these systems evolve, they may increasingly override critical thinking processes and erode individuals' capacity to distinguish between fact and manipulation.

Concluding remarks

This article highlights the urgent need for cross-organizational, cross-sector and international collaboration and open communication in tackling disinformation and the imminent risks AI-augmented disinformation poses to our democratic societies.

What is certain is that a unified, proactive, and forward-looking strategy on AI and disinformation is urgently needed, both at the national and EU levels. Current approaches to AI policymaking often create strategic vulnerabilities, either due to their reactive and fragmented nature or because policy decisions are not effectively communicated to all relevant stakeholders. With this article, we hope to encourage our participating organizations to deepen their collaboration and collective efforts.

While the article primarily focuses on risks and critical issues, its message is not without hope. AI tools also offer opportunities that can be harnessed to counter the growing information disorder and strengthen societal resilience. Additionally, there are a number of other emerging tools and initiatives that provide hopeful avenues. For example, Faktabaari's initiative demonstrates how educational effort focused on AI can contribute to societal resilience. By targeting educators as multipliers, their AI Guide for Teachers, aims to equip future generations with the critical skills needed to navigate an AI-saturated information environment.