Think7 ITALY 2024

G7 ITALIA 2024

Policy Brief

May 2024

Task Force 1: Peace, Security and Global Governance

Mitigating AI-Generated Disinformation: A Cyber Collaborative Framework for G7 Governance

Leonardo De Agostini, CFI Project Officer, EU Institute for Security Studies (EUISS), France
Beatrice Catena, CFI Project Officer, EU Institute for Security Studies (EUISS), France
Simona Autolitano, Center for Advanced Security, Strategic and Integration Studies (CASSIS), Germany

Abstract

In an increasingly multipolar and conflict-prone world, witnessing the rise of artificial intelligence (AI), this paper explores the disruptive potential of AI-generated disinformation, a growing threat to global peace and security. The advance of generative AI tools able to rapidly produce convincing "synthetic disinformation", such as large language models (LLMs), has exponentially amplified the reach and impact of foreign information manipulation and interference (FIMI) wielded by both state actors and non-state actors alike. Against this challenge, the G7 countries have shown proactive leadership in recognising and addressing the threat posed by AI and disinformation, leveraging resources and expertise to develop innovative strategies. Still, the lack of uniformity in regulatory approaches and policies across G7 nations, as well as the compartmentalisation between cyber policies and counter-disinformation responses, has resulted in fragmented solutions that are now insufficient. This paper argues that tackling AI-generated disinformation with a cyber-security approach not only offers an effective framework for G7 action but also paves the way for broader AI and cyber regulation milestones, leveraging the G7's role as norm setter in peace, security and global governance.

1. The geopolitics of AI: Implications for the G7

The year 2024 saw the inevitable clash of two emerging realities that were long destined to meet, with global repercussions as a consequence: an increasingly connected, multipolar and conflict-ripe world and the emergence of artificial intelligence (AI). The interlinks between these two realities and the signs of this clash's disruptive potential are evidenced by the escalation of cyber threats and AI-generated disinformation campaigns. If the Russian war of aggression in Ukraine and the Israel-Hamas conflict show some early examples, the full potential of AI generated disinformation and AI powered cyber-attacks is yet to manifest. Given their paramount role as norm-setters in democratic governance, their positioning at the forefront of geopolitical competition, and the home of numerous so-called very large online platforms (VLOPs), G7 nations find themselves at the heart of this nexus.

If generative AI "broke into the public consciousness" in 2022 (Maslej et al. 2023: 90), its potential as a multiplier of capabilities for malicious state and non-state actors alike was finally identified in this year's World Economic Forum (2024) Global Risks Report. In parallel the rapidity with which synthetic content – ranging from images and videos to voice cloning, reached a high degree of realism has been included in this year's Munich Security Report (Bunde et al. 2024).

The detrimental impact of the misuse of generative AI on peace and security is evident in two distinct contexts: its manipulation of democratic processes within G7 nations and its interference with their foreign policy endeavours, particularly in times of conflict.

In the case of the former, the expected impact of an unregulated generative AI on democratic processes and elections is undisputed. Studies proved that generative AI tools are accepting with varying but overwhelmingly positive degrees prompts trying to generate electoral disinformation (Walter 2023; CCDH 2024). The potential of large language models (LLMs) to produce convincing English language content (Goldstein et al. 2024) – especially if coupled with minimum human supervision – is growing exponentially. With LLMs rapidly improving, it is clear how they could be used to 'supercharge' both foreign and domestic information manipulation.

At the same time the world is witnessing a resurgence of information warfare and influence operations, where generative AI is increasingly part of the mix on both fronts: advancing offensive capabilities while potentially optimising countermeasures like detection (Bozalka 2023; Fredheim and Pamment 2024). In the foreign policy realm, this has led to the formulation of concepts as foreign information manipulation and interference (FIMI) to describe a manipulative and intentional pattern of behaviour that “threatens or has the potential to negatively impact values, procedures, and political processes” (EEAS 2023: 25). Indeed, the current geopolitical context offers numerous examples.

It took just a couple of weeks after the Russian aggression on Ukraine for a deepfake video of Ukrainian President Zelensky to appear online calling for his soldiers to surrender.[1] While the “algorithmically-driven fog of war”[2] – characterised by gruesome deepfakes of war crimes and atrocities in turn attributed to both warring parties,[3] characterises the Israel-Hamas conflict. If it was argued that the effects of generative AI on the information landscape are exaggerated (Simon et al. 2023), these certainly spark a deeper conversation regarding the power of AI to mislead transnational public opinions. “What happens when literally everything you see that’s digital could be synthetic?”[4]

If the ‘geopolitics of AI’ has been under the spotlight and preoccupations of policymakers in the past, the exponential acceleration of AI capabilities and its increasingly widespread availability have prompted states around the globe to engage in a real race to control the future of AI (Smuha 2021). Against the backdrop of accelerated geopolitical tensions and rising global norm-contestation, this race is not just technological: it is at heart a regulatory race.

In this context, G7 members have a fundamental role to play and have already taken significant steps. Indeed, AI has been under the spotlight of previous G7 presidencies, most notably with the Hiroshima Process on establishing the International Code of Conduct for Organizations Developing Advanced AI Systems. Here, AI’s potential misuse as a disinformation tool is addressed in regard to the need for “reliable content authentication and provenance mechanisms” which would allow for “users to identify AI-generated content” (G7 2023: 6).

More recently, two important strategic documents were issued on the two sides of the Atlantic: the United States Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence and the European Union AI Act. The first one tackles the problem of possible societal harm caused by AI-generated disinformation stressing the importance of labelling synthetic content, while

[1] See for example Wakefield 2022.

[2] As defined by Avi Asher-Schapiro, from the Thomson Reuters Foundation. See Gladstone 2023.

[3] See for example Klepper 2023.

[4] William Marcellino, senior behavioural and social scientist at the RAND Graduate School quoted in the New York Times. See Hsu and Thompson 2023.

the latter stresses that generative AI applications would similarly need to respect transparency requirements – such as disclosing which content is AI-generated.

If the possibilities offered by the AI revolution will span over all sectors of government, the growing use of disinformation campaigns to destabilise societies and democratic processes offers a compelling example of this technology's impact on peace and security, while also showcasing how an effective cyber governance could safeguard an open and democratic international system.

This policy brief contends that the G7 holds a distinct advantage in leading the charge against AI-generated disinformation. By spearheading efforts to develop tangible policy adjustments, the G7 can effectively address this pressing issue and pave the way for a more secure and resilient cyberspace. Through collaborative endeavours and strategic initiatives, the G7 can capitalize on its position to enact comprehensive regulatory measures that mitigate the threats posed by AI-enabled disinformation.

2. Understanding the disrupting power of AI-generated disinformation

The rise of AI-generated disinformation is having a dramatic impact on elections around the world, with examples ranging from Moldova to Taiwan. Examples of AI deepfakes influencing electoral processes encompass a broad array of events, including a video depicting the Moldovan president endorsing a pro-Russian party, audio recordings featuring a Slovak party leader discussing vote rigging, and a video portraying an opposition lawmaker in Bangladesh in a compromising situation.

While it can be argued that disinformation is not a new phenomenon and has existed for centuries, the internet and advances in AI technology have made it even more effective. This has resulted in greater reach and believability of fake news. In particular, the emergence of generative AI, and specifically the development of LLMs and their role in the creation of deepfakes represents a significant shift in the AI landscape.

Previously, the creation of convincing dialogue for deepfake videos necessitated manual composition, time and skills. Today, LLMs have streamlined this process, allowing individuals to delegate dialogue generation to AI systems such as ChatGPT or Microsoft's Bing chatbot. By providing a mere outline of the content, users can effortlessly procure authentic-sounding dialogue, thereby minimising time and effort (Mitra et al. 2024).

LLMs differ from previous AI models on various fronts. Firstly, they are trained using an immensely large dataset. Secondly, they generate a human-like output, which is immediately understandable and usable in real-world scenarios, such as in the form of text, unlike previous models that generated labels and categories. Thirdly, they have a broad scope and considerable autonomy in extracting patterns from large datasets, enabling them to generate new content. Fourthly, multimodal systems are capable of processing multiple types of inputs simultaneously, including text, images, and audio.

The application of these characteristics to the production of AI-generated disinformation poses a significant challenge. Indeed, LLMs have the ability to generate more convincing disinformation and to disseminate and produce disinformation more effectively due to their scale, speed, and ease of use. The use of LLMs can lead to the automated mass production of fake news, making it difficult to detect and moderate. False information can be generated rapidly, in just a few seconds, outpacing conventional fact-checking procedures. Furthermore, the accessibility and user-friendliness of AI tools have reduced the barriers to entry, making them widely available to a large number of users without requiring specialised expertise (Bashardoust et al. 2024).

LLMs have the potential to benefit society in various fields of application. However, they also present a challenge to current societies as they can facilitate the production of fake news, whether intentionally or unintentionally.

On the one hand, as already highlighted, their harmful applications involve the creation of false content for diverse motives, ranging from financial gain to political agendas. LLMs enable the generation of AI-generated content that is highly targeted and personalised, with the resulting content being almost indistinguishable from content generated by humans. This enables fraudulent activities such as scams, phishing campaigns, and cyberattacks.

Furthermore, AI models can also be intentionally attacked with the objective of deceiving the model and inducing incorrect outputs. For example, AI-based tools used to combat disinformation can be targeted to bypass restrictions and increase the dissemination of AI-generated disinformation.

On the other hand, LLMs can unintentionally produce disinformation by generating plausible text that diverges from data inputs and lacks a verifiable basis. This phenomenon is known as "hallucination" and can facilitate the generation and spread of fake news and biased results.

AI-enabled disinformation, whether intentional or not, poses a significant threat to the G7 nations. It has the capacity to manipulate public opinion, undermine democratic processes, and sow social discord.

However, the current approach to countering AI-generated disinformation is not yet fully fledged. There is too much compartmentalisation between the cybersecurity and disinformation communities.

The reality is that AI-generated disinformation is not typically considered a cybersecurity threat in the conventional sense. While cybersecurity has traditionally focused on protecting computer systems and digital infrastructure, disinformation exploits human vulnerabilities such as cognitive biases and logical fallacies. Indeed, despite their apparent differences, there is a significant overlap between the tactics used by cyber attackers and those who spread disinformation. Adversaries often use a mixture of cyber-attacks and disinformation tactics to achieve their objectives.

However, the prevailing government approach has been to treat these phenomena as distinct, resulting in separate communities and strategies to counter each. This compartmentalisation overlooks the interconnectedness of cyber threats and disinformation campaigns, thus resulting in ineffective countermeasures.

To better explain this lack of consideration of AI-generated disinformation, we can look to most G7 nations, such as Germany, which defines disinformation as a 'hybrid threat'. The Federal Ministry of the Interior (BMI) coordinates the federal government's approach to hybrid threats – meaning it coordinates a number of agencies and bodies working in this area – but there is no clear vision of who is primarily responsible for tackling these challenges, with the German cybersecurity agency, the Federal Office for Information Security (BSI), playing a very limited role.

The same is true for the EU as a whole, where despite significant progress in EU activities to regulate AI, EU legislation lacks specific rules for AI-generated disinformation and there is fragmentation between cybersecurity and disinformation measures (Novelli et al. 2024).

Furthermore, the benefits of collaborating with the private sector are often underestimated. While private companies are at the forefront of AI technology development, they are frequently overlooked as part of the solution.

Many companies, regardless of their size, are already investing substantial resources in researching ways to counter AI-generated disinformation. However, there is a noticeable lack of coordination with governments in this effort, hindering the effectiveness of current strategies. Public-private partnerships (PPPs) can bridge this gap. Just as in the cybersecurity world, these partnerships can support public-private collaboration in finding innovative tools to counter AI-generated disinformation. By pooling resources, expertise, and networks, PPPs can enhance information sharing, coordination and improve the implementation of countermeasures. Examples like Germany's Alliance for Cybersecurity,[5] offer valuable insights for effective collaboration, serving as a model for countering AI-generated disinformation.

While all G7 countries have great expertise and cybersecurity agencies doing great work in the field of AI that can be put at the service of society, integration with the disinformation community is more sporadic or non-existent.

[5] For more information on this initiative of the Federal Office for Information Secuirty (BSI) see the Alliance for Cybersecurity website: https://www.allianz-fuer-cybersicherheit.de.

3. Building synergies: Integrating efforts for enhanced cyber governance

In part, the issue of strategic and conceptual compartmentalization of this threat contributes to a fragmented regulatory approach (Schmitt, 2021), and an excessive stratification of actions. Over the past decade, the expansion of initiatives undertaken by G7 nations to counter disinformation and hybrid threats, including AI-enabled disinformation, is noteworthy (Juršėnas 2022). At large, these efforts against the spread of AI-enabled disinformation can be categorized into three spheres of action, starting with a strategic commitment to countering the threat (1), proactive/comprehensive policies to tackle its risks (2), and more tailored threat-specific initiatives (3).

First of all, whether in the national or international arena, all G7 nations have made significant strategic commitments to countering disinformation by implementing comprehensive frameworks that address its impact on democracy and society, while proposing solutions. This dedication extends to their participation in multilateral frameworks such as the Hybrid Centre of Excellence (Hybrid CoE), where all G7 states, except Japan, are active members. Through their engagement in Hybrid CoE initiatives, these countries affirm a shared understanding of hybrid threats, encompassing cyber and AI-enabled disinformation within a broader context (Mazzucchi 2023).

Second, G7 nations like Canada and the United Kingdom have taken proactive stances against AI-enabled disinformation, focusing on early detection and intervention to prevent its escalation (2). For instance, Canada's Critical Election Incident Public Protocol (CEIPP), launched in 2019 (Canada 2023), fosters collaboration and information sharing among stakeholders in the lead-up to elections, while reiterating Canada's commitment to the G7 Rapid Response Mechanism (RRM). Similarly, the UK's Online Safety Act, passed in October 2023, exemplifies this approach by aiming to protect users from harmful content and holding tech companies accountable.

Third, targeted issue-specific regulations have been implemented to address specific aspects of AI-generated disinformation across all G7 nations (3). Here, France provides a notable example. While on one hand, the country has actively established a Task Force dedicated to countering disinformation in elections since 2017, it has also spearheaded case-specific research through its VIGINUM agency, tasked with uncovering fake content and disinformation (VIGINUM 2024).

Undoubtedly, this extensive range of measures taken by G7 nations, spanning national, multilateral, and targeted efforts, underscores the collective recognition of the seriousness of the disinformation threat, albeit not specifically tailored to the nuances of AI-generated disinformation. G7 nations have yet to adequately confront the intricate technical challenges inherent in thwarting the weaponization of AI by malicious actors. In part, the diverse strategic cultures among G7 nations, their history of collaboration with the private sector, and their foreign policy priorities directly or indirectly shape each nation's approach to this threat.

Still, and while the dispersed nature of initiatives across the G7 is a challenge, it also presents distinct opportunities. The renewed strategic relevance of disinformation and hybrid threats more broadly has equipped G7 nations with a pool of knowledge and expertise unprecedented in scale. While this is true for the hybrid field, it is even more relevant in the cyberspace. Furthermore, the (at times fuzzy) coexistence of both national and international efforts has proven that even though national sovereignty remains key, G7 nations are willing and capable of leveraging cooperation in this space.

Hence, the G7 holds a unique advantage, as well as the capability, to spearhead innovative solutions in the ongoing battle against AI-enabled disinformation. By bridging the conceptual understanding gap regarding AI-enabled disinformation as a cyber threat and rectifying regulatory deficiencies, the G7 can pioneer initiatives that target the root causes of the issue, rather than merely treating its symptoms with a patchwork of measures.

This collective endeavour holds the potential to elevate global efforts in combating disinformation to unprecedented levels of efficacy and impact, promising a new frontier in the ongoing battle for cyber governance and peace.

4. Towards shared solutions: G7 policy recommendations and the path forward

The scattered hybrid and cyber responses have not effectively countered AI-generated disinformation. Nonetheless, the G7's considerable expertise and resources can serve as a foundation for addressing this issue. This can be rectified in two ways:

First, the G7 can serve as a platform to overcome the gap between cyber and hybrid solutions to disinformation, as this is inherently bridged by the risks posed by AI-enabled disinformation. In practice, this means:

- *Leveraging cyber expertise*: the response to disinformation remains fragmented and lacks the coordination that is critical to success. Cybersecurity initiatives have invested heavily in security best practices, establishing robust frameworks, guidelines and standards. These efforts also emphasized fostering collaboration between the public and private sector through initiatives such as PPPs led by governments, the use of threat modelling and the maintenance of global databases of vulnerabilities and known bugs. These insights could serve as a model for addressing the challenges posed by AI-generated disinformation. Moreover, the response to disinformation should emulate the defence-in-depth strategy of cybersecurity, ensuring a layered approach with multiple defences in place. This includes a continuum of human and AI monitors that verify authenticity and fact-check, intervening before false information is disseminated or removing it afterwards.

- *Leveraging cyber regulations*: while both the EU AI Act and the US Executive Order acknowledge the hybrid risks associated with the threat of AI-enabled disinformation, neither considers the risks associated with its misuse. The legislative nature of these regulations offers an unprecedented space for action that should be harnessed by G7 nations.
- *Establishing a permanent G7 Working Group for National Cybersecurity Agencies*: as recognized by the Group, their vital role in addressing malicious cyber activities and FIMI could foster international collaboration at a technical and political level. The working group should be established as permanent to ensure continuation.

Second, the G7 should promote international initiatives that foster growth in expertise on AI-generated content, monitoring capability and building resilience. This includes:

- *Promoting research initiatives focused on detecting AI-generated content*: sharing insights and tools for identifying AI-generated content among allied governments via a dedicated intergovernmental framework, collaborating on exchanging details regarding specific incidents, their repercussions, and strategies for handling them.
- *Building resilience across societies*: education is proving to be a powerful tool in mitigating cybersecurity risks and building resilience, with both employees and the public receiving training to raise awareness of threats such as phishing emails and malware. Similar efforts must be made to educate individuals on how to recognise and combat disinformation, with a holistic whole of society approach.
- *Leading international norm setting for content authentication and provenance mechanisms*: in line with the principles agreed in the Hiroshima Process, G7 members could identify a coordinated mechanism to share best practices and evaluate the organisations' efforts in line with the risk-based approach identified in the Code of Conduct. For example, a G7 action encouraging standard setting for AI-generated content labelling.
- *Invest in the RRM*: consider how to integrate the new challenges arising from AI-generated disinformation into the existing coordination on evolving foreign threats to democracy via the existing G7 RRM.

References

Bashardoust, Amirsiavosh, Feuerriegel, Stefan, and Shrestha, Yash Raj. 2024. Comparing the willingness to share for human-generated vs. AI-generated fake news. *arXiv* 12 February. https://doi.org/10.48550/arXiv.2402.07395

Bozalka, Dusan. 2023. Information warfare in the age of artificial intelligence. *IRSEM Strategic Briefs* 62. https://www.irsem.fr/publications-de-l-irsem/breves-strategiques/strategic-brief-no-62-2023-information-warfare-in-the-age-of-artificial-intelligence.html

Bunde, Tobias, Eisentraut, Sophie, and Schütte, Leonard, eds. 2024. *Munich security report 2024: Lose-lose?*. Munich: Munich Security Conference. https://doi.org/10.47342/BMQK9457

Canada Government. 2023. Foreign interference: Critical election incident public protocol. *Parliamentary committee notes*. https://www.publicsafety.gc.ca/cnt/trnsprnc/brfng-mtrls/prlmntry-bndrs/20230629/07-en.aspx

CCDH-Center for Countering Digital Hate. 2024. *Fake images factories: How AI image generators threaten election integrity and democracy*. https://counterhate.com/?p=3749

EEAS-European External Action Service. 2023. *1st EEAS report on foreign information manipulation and interference threats*. https://www.eeas.europa.eu/node/425201_en

Fredheim, Rolf, and Pamment, James. 2024. Assessing the risks and opportunities posed by AI-enhanced influence operations on social media. *Place branding and public diplomacy* 8 February. https://doi.org/10.1057/s41254-023-00322-5

G7. 2023. *Hiroshima process international code of conduct for organizations developing advanced AI systems*. http://www.g7.utoronto.ca/summit/2023hiroshima/231030-ai-code-of-conduct.html

Gladstone, Brooke. 2023. How Elon Musk's X failed during the Israel-Hamas conflict. *On the Media* 11 October. https://www.wnycstudios.org/podcasts/otm/episodes/on-the-media-how-elon-musks-x-failed-during-israel-hamas-conflict?tab=transcript

Goldstein, Josh A., et al. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus* 3(2): pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Hsu, Tiffany and Thompson, Stuart A. 2023. A.I. muddies Israel-Hamas War in unexpected way. *The New York Times* 28 October. https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamas-war-in-unexpected-way.html

Juršėnas, Alfonsas, et al. 2022. *The role of AI in the battle against disinformation*. Riga: NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/the-role-of-ai-in-the-battle-against-disinformation/238

Klepper, David. 2023. Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead. *AP News* 28 November. https://apnews.com/article/a1bb303b637ffbbb9cbc3aa1e000db47

Maslej, Nestor, et al. 2023. Artificial intelligence index report 2023. *arXiv* 5 October. https://doi.org/10.48550/arXiv.2310.03715

Mazzucchi, Nicolas. 2022. AI-based technologies in hybrid conflict: The future of influence operations. *Hybrid CoE Papers* 14. https://www.hybridcoe.fi/publications/hybrid-coe-paper-14-ai-based-technologies-in-hybrid-conflict-the-future-of-influence-operations

Mitra, Alakananda, Mohanty, Saraju P., and Kougianos, Elias. 2024. The world of generative AI: Deepfakes and large language models. *arXiv* 6 February. https://doi.org/10.48550/arXiv.2402.04373

Novelli, Claudio, et al. 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *arXiv* 15 March. https://doi.org/10.48550/arXiv.2401.07348

Simon, Felix M., Altay, Sacha, and Mercier, Hugo. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review* 4(5). https://doi.org/10.37016/mr-2020-127

Sindermann, Cornelia, et al. (2021). The evaluation of fake and true news: on the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption. *Heliyon* 7(3): e06503. https://doi.org/10.1016/j.heliyon.2021.e06503

Smuha, Nathalie A. (2021). From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology* 13(1): 57-84, https://doi.org/10.1080/17579961.2021.1898300

Swenson, Ali, and Chan, Kelvin. 2024. Election disinformation takes a big leap with AI being used to deceive worldwide, *AP News* 14 March. https://apnews.com/article/bc283e7426402f0b4baa7df280a4c3fd

VIGINUM. (2024). *Portal Kombat: a structured and coordinated pro-Russian propaganda network*. https://www.sgdsn.gouv.fr/publications/portal-kombat-un-reseau-structure-et-coordonne-de-propagande-prorusse

Wakefield, Jane. 2022. Deepfake presidents used in Russia-Ukraine war. *BBC News* 18 March. https://www.bbc.com/news/technology-60780142

Walter, Kyle. 2023. Testing multimodal generative AI: Generating elections mis-and-disinformation evidence. *Logically Reports*. https://www.logically.ai/resources/generative-ai

World Economic Forum. 2024. *Global Risks Report 2024*. https://www.weforum.org/publications/global-risks-report-2024/digest

About Think7

Think7 (T7) is the official think tank engagement group of the Group of 7 (G7). It provides research-based policy recommendations for G7 countries and partners. The Istituto Affari Internazionali (IAI) and Istituto per gli Studi di Politica Internazionale (ISPI) are the co-chairs of T7 under Italy's 2024 G7 presidency.