PREMIER MINISTRE

Liberté
Égalité
Fraternité

Secrétariat général de la défense et de la sécurité nationale

Challenges and opportunities of artificial intelligence in the fight against information manipulation

VIGINUM

Systemic issues

AI ACTION SUMMIT

This report was produced by VIGINUM, with international contributions from the European External Action Service (EEAS), the Psychological Defence Agency (Sweden), Canada's Rapid Response Mechanism (RRM Canada), the Foreign and Commonwealth Office (UK), and two independent fact-checking organisations from civil society: *Lupa* (Brazil) and *Full Fact* (UK).

As part of its remit under Article 3 of Decree No. 2021-922 of 13 July 2021, VIGINUM (France's service for vigilance and protection against foreign digital interference) is responsible for detecting and characterising foreign digital interference operations by analysing the content publicly accessible on online platforms. To this end, VIGINUM is authorised by Decree No. 2021-1587 of 7 December 2021 to carry out automated processing of personal data.

Foreign digital interference, the digital component of information manipulation, is an operation "*involving, directly or indirectly, a foreign State or a foreign non-State entity, aimed at the artificial or automated, massive, and deliberate dissemination, via an online public communication service, of allegations or insinuations of facts that are manifestly inaccurate or misleading and likely to harm the fundamental interests of the Nation* ".

Since the early 2020s, there has been unprecedented growth in artificial intelligence (AI) applications, particularly generative AI (GenAI). Indeed, the ergonomics of products as well as their easy, free access have encouraged their appropriation by a wider audience: thus, according to BPI France,[1] by 2027, there could be half a billion regular users of AI-related technologies worldwide.

While certain developments hint at promising prospects in the areas of health, the environment, and mobility, the widespread use of generative tools nevertheless raises questions about their impact on the integrity of the information environment. In particular, it raises fears of a structural increase in the threat level associated with foreign digital interference, with the risk of altering citizens' perception of reality. Indeed, with the spread of AI tools, and consequently the likely profusion of artificially generated content on online platforms, it will become more difficult to distinguish genuine from synthetic. When it comes to information, the consequences of a distorted reality are likely to produce lasting and negative effects on the functioning of democracy, particularly during elections, and more generally lead to the destabilisation of our societies, which are founded on the notion of trust.

In an international context marked by the balance of power and interdependence, in which the mastery of technology is asserting itself as a major factor of power, it is useful to consider the actual consequences of AI on the information threat, while also exploring its use cases in the fight against information manipulation.

Beyond the promises and fears associated with the use of AI, this report aims to take a realistic look at the technological challenges: while AI technologies are likely to increase certain capabilities of the information threat, AI also offers real opportunities to strengthen our defences against information manoeuvres, whatever their degree of sophistication.

This report, presenting the combined viewpoints of VIGINUM and international players, both institutional and from civil society, has three main objectives: **to contribute to public information and awareness,** by raising the level of knowledge of cases of malicious use of AI for the purposes of information manipulation; to **highlight the opportunities offered by AI** in the fight against information manipulation, in order to promote the international sharing of best practices ; and finally, to **encourage cooperation** between institutional players, civil society, the academic and scientific world, and private actors, in order to speed up the development of innovative solutions for the benefit of the entire ecosystem responsible for fighting information manipulation.

[1] Source: BPI France le Hub, based on Statista data, see https://bigmedia.bpifrance.fr/nos-actualites/marche-de-lintelligence-artificielle-ou-en-sommes-nous; and https://comarketing-news.fr/le-marche-de-lia-va-doubler-en-4-ans/

I. AI, a technology used by information threat actors

1. Definitions

Artificial intelligence (AI) refers to all logical and automated processes, generally based on algorithms, designed to reproduce, at least in part, human behaviour, such as learning, reasoning, planning, or creating.

Artificial intelligence is based on three main concepts:

- The aim of machine learning is to give machines the ability to learn autonomously, using mathematical models, in order to extract the most relevant information from a set of training data made available to them;
- A sub-category of machine learning, deep learning is an automatic learning process using artificial neural networks to solve complex problems such as pattern recognition (artificial vision) or natural language processing;
- Finally, generative AI (GenAI) technologies, which are a sub-segment of deep learning and can be used to generate text, images or computer code, represent a change of scale in content generation. These technologies made a major breakthrough in 2022, when web services were made available to the general public, enabling text and image content to be generated very easily.

2. An attractive technology for information threat actors

NB: The trends and use cases documented in parts 2) and 3) cover only the operating methods using generative AI, observed in the context of open-source investigations. The potential uses of other AI segments in the context of information manoeuvres (e.g., use of machine learning algorithms to target specific audiences) are not covered here, as these are not easily documented in open sources.

In the context of the now widespread use of GenAI technologies in many fields, over the last two years VIGINUM has observed an increasing use of GenAI by the various foreign players in the field of information threat.

Three main trends characterise the impact of this technology on today's information threat landscape:

- A change of scale in the generation of potentially inaccurate or misleading content;
- Tenfold capacity increases in large-scale replication and coordinated publication of inauthentic content;
- A tool for generating and managing inauthentic accounts on online platforms.

a. A change of scale in the generation of potentially inaccurate or misleading content

Like the new capabilities offered by AI for many business sectors (health, transport, environment), the use of GenAI technologies is likely to significantly increase the ability of malicious actors to generate manifestly inaccurate or misleading content in the information space, mainly through three types of use.

First of all, these technologies enable **them to optimise their basic productivity** (translation, grammatical and spelling correction, document synthesis, etc.).

They also enable the **large-scale generation of texts in various forms**. These can be both long texts (e.g., articles), short texts (posts for social networks, comments, etc.), texts aimed at a wide audience, or tailor-

made texts targeting very specific audiences (through prompts[2] designed to imitate the tone and language of personas representative of certain categories of the population for the purposes of *microtargeting*). This use covers varying levels of sophistication, from simple text generation to more complex uses (running networks of accounts via GenAI, automating the analysis of publications and generating responses in return, based on AI agents[3]). The use of GenAI to generate text content is theoretically likely to facilitate the deployment of sophisticated information manoeuvres, by allowing greater diversity in the rewording of the same text or narrative, as well as disrupting the identification of the inauthentic, coordinated nature of a campaign. This avoids the pitfalls of copy-pasta,[4] which is easily detected and penalised by the platforms.

Finally, these technologies make it possible to **simplify the generation of visual, audio, or video content**, with a level of stylisation and quality that is likely to attract attention and generate engagement. Although much of the synthetic content disseminated on social networks is evidence of the recreational use of AI, in the register of humour or irony,[5] VIGINUM's observations also point to a growing use of these technologies for malicious information manoeuvres (see section I.3 below).

These productivity gains in terms of content generation also enable players in the information threat arena to develop their strategies over the medium and long term. For example, *Newsguard* has identified 1,150 news websites that are generated entirely or mainly by AI language models in a total of 16 different languages.[6] In addition to the fact that such sites are pre-positioned devices that can be activated as part of information manoeuvres, they also enable their operators to capture large volumes of programmatic advertising spend[7] to the detriment of genuine media, using a variety of methods ("clickbait" articles, auto-play videos, contextual advertising, etc.).

b. Tenfold capacity increases in large-scale replication and coordinated publication of inauthentic content

The ease with which content can be created using GenAI technologies thus increases tenfold the ability of malicious actors to carry out large-scale coordinated actions on several platforms, and even helps to internationalise their strategies. In addition to the ability to automatically produce different versions of the same text with slight rewordings to fool detection mechanisms, AI technologies also make it easier to write content in many languages, in a convincing syntax, making it easier to target new countries or to target certain linguistic communities (e.g., diasporas) more specifically within the same country.

c. A tool for generating and managing inauthentic accounts on online platforms

One of the technical challenges faced by those involved in information manipulation is optimising the distribution of content on online platforms. Indeed, a foreign digital interference operation needs vectors to reach its target audience. However, managing and running the inauthentic accounts used for distribution raises issues of profile credibility, stealth, and anonymity of the activity.

---

[2] A prompt is an instruction or set of data provided to an AI system, which uses this information to produce responses or creations in the form of text, images, or other types of media.
[3] https://huggingface.co/blog/ethics-soc-7
[4] A block of text or visual material copied and pasted identically or almost identically onto one or more web platforms in order to increase the visibility of a message.
[5]https://www.francetvinfo.fr/replay-radio/le-vrai-du-faux/macron-en-eboueur-trump-interpelle-obama-et-merkel-a-la-plage-les-images-creees-par-des-intelligences-artificielles-sont-de-plus-en-plus-realistes_5712170.html
[6] Newsguard AI Tracking Center, consulted on 7 January 2025, https://www.newsguardtech.com/special-reports/ai-tracking-center/
[7] Automated process for buying and selling digital advertising space.

As pointed out in a report by *OpenAI*,[8] the use of AI-generated images is widely employed in information manipulation campaigns, to provide fake accounts with a credible profile picture, but also to industrialise and give credibility to the generation of inauthentic accounts: name, biography, age, content, etc.

Therefore, a combination of text generation models and visual generative models is likely to greatly facilitate the creation of inauthentic accounts, despite the inauthentic behaviour detection systems put in place by the platforms themselves.

3. Observed uses of GenAI in foreign digital interference

Drawing on the work of VIGINUM and leading international actors, this section aims to document various observed cases of GenAI used for foreign digital interference, based on open-source investigations.

a. Main documented operating methods

As part of its remit and since 2022, VIGINUM has observed a growing use of generative AI by foreign threat actors targeting the French-speaking digital public debate.

i. Text generation

This method involves the large-scale production of text content using GenAI applications. By way of example, VIGINUM observed on several occasions during 2024, particularly in the context of the Paris Olympic and Paralympic Games, the dissemination of texts most likely generated by AI[9] on various platforms such as *X* and *Tumblr*, by ecosystems of accounts with inauthentic characteristics, most likely affiliated with the *Spamouflage* operating method[10]. Although these manoeuvres were massive in terms of the volume of messages posted, in the end they generated very little engagement, mainly because the inauthentic posting accounts had only a small audience, thereby mitigating the risk of impact of such coordinated posts.

Examples of posts by a network of inauthentic accounts that distributed AI-generated text content

Similarly, in 2024, security firm *CyberCX*[11] observed a network of accounts on *X* with inauthentic characteristics whose main activity was to amplify divisive political topics using generative AI as their predominant operating method. Posts from this network, identified by the very phrase "*As an AI language*

[8] https://openai.com/global-affairs/an-update-on-disrupting-deceptive-uses-of-ai/.
[9] As with the two post examples mentioned above, VIGINUM identified a total of around one hundred messages disseminating an almost identical narrative during this manoeuvre.
[10] The *Spamouflage* modus operandi, documented publicly for the first time by *Graphika* in 2019, is a pro-PRC influence mechanism that aims to disseminate narratives favourable to the interests of the Chinese Communist Party (CCP) to an international audience. It is based on networks of accounts with inauthentic characteristics, responsible for carrying out information manoeuvres on a multitude of platforms. See https://www.graphika.com/reports/spamouflage.
[11] See https://connect.cybercx.com.au/Intelligence-Update-CCX-IU-2024-004.

model..." at the beginning of the message, were not very visible and again, generated very little engagement.

In addition, as mentioned above, AI text generation can also be used to supply fake online news media sites by creating numerous articles with blatantly inaccurate or misleading content. In this respect, several sites linked to the *RRN* ecosystem[12] are stocked with articles generated by AI, which are then distributed to French audiences via targeted online advertising. The full range of possibilities offered in terms of text generation also enables such threat actors to optimise their content translation capabilities, which in turn enables them to improve the search engine optimization for such fake news sites.

ii. Image generation

VIGINUM has also observed a growing use of AI-generated images by information threat actors. Unlike text generation, the purpose of using artificially generated images appears mainly to be to illustrate a given narrative with striking or symbolic images, rather than to genuinely deceive users concerning their authenticity.

Such content, which may mimic a cartoonish style, use bright colours or visual elements likely to create emotion, seeks to capture the attention of Internet users within their news feed, in order to support a given narrative with a view to generating engagement.

In particular, VIGINUM observed the distribution of AI-generated visuals by accounts with inauthentic characteristics affiliated with the *Spamouflage* operating method. These images are mostly drawings and caricatures, with a few realistic representations of well-known sites.

Screenshots of AI-generated visuals posted on X and Tumblr by Spamouflage accounts.

On *X*, certain ultra-conservative[13] trolling accounts regularly broadcast images generated by AI, as reported by the Swedish public television company *SVT* in a recent investigation[14] carried out on two accounts, one of which has several hundred thousand followers. In the case in point, this operating method enables trolling accounts to support polarising narratives (e.g., anti-immigration) and generate several million views on *X*, at a lower cost.

[12] https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et.

[13] Trolling is online behaviour in which an individual seeks to create tension, provoke, or divert a conversation from its initial objective with offensive or completely off-topic messages.

[14] https://www.svt.se/nyheter/utrikes/sa-gjorde-vi-granskningen-av-europe-invasion.

iii. Video and audio generation

This method involves producing synthetic video or audio content that is clearly inaccurate or misleading, in order to distribute it on several platforms. The synthetic content generated may be original, or it may offer a modified version of authentic content, by altering one or more of its elements (voice, face, appearance, etc.).

While synthetic videos appear to be used increasingly on social networks to entertain partisan audiences, the use of credible synthetic videos as part of information manoeuvres to mislead users seems marginal for the moment. The cost of producing such content (mixing multiple AI technologies, video editing, subtitle overlay, etc.) and the fact that it is probably less easy to distribute on social networks make this method less accessible to malicious actors.

However, on 13 February 2024, *France 24*'s investigative unit, "Les Observateurs", detected and denounced the broadcasting of a video manipulated by AI, usurping the identity of its journalist Julien FANCIULLI. In the video, the journalist claimed that Ukrainian intelligence had planned to assassinate Emmanuel MACRON and to blame Russia for it, in order to obtain new arms deliveries.

Screenshot of the fake France 24 *report on* Telegram.

After analysis, "Les Observateurs" found that "the presenter's lip movements were not synchronised with the words spoken in the video [...] the vocal intonation seemed robotic" and that "certain formulations [...] did not correspond to the way in which information is given on air".[15]

Furthermore, on 13 December 2024, the independent Russian-language investigative media *The Insider* published an article presenting a new *Matryoshka* campaign (documented by VIGINUM last June[16]) aimed at convincing Internet users that professors from prestigious universities were calling on the West to lift sanctions against Russia, while criticising Ukrainian President Volodymyr ZELENSKY. Investigations showed the use of AI tools in these videos, in particular to clone the voices of academics, some of whom had confirmed that the statements were not made by them.[17]

Screenshot of a video disseminated by the Matryoshka *operating method.*

[15]https://observers.france24.com/fr/%C3%A9missions/les-observateurs/20240214-une-tentative-d-assassinat-contre-emmanuel-macron-en-ukraine-attention-cette-vid%C3%A9o-est-truqu%C3%A9e.
[16] https://www.sgdsn.gouv.fr/publications/matriochka-une-campagne-prorusse-ciblant-les-medias-et-la-communaute-des-fact-checkers.
[17] https://theins.press/en/news/277174.

b. Various views on the information threat using GenAI: international institutional players

In order to provide a broader overview of these issues, this section includes the perspective of leading international players in the malicious use of AI for foreign digital interference.

i. European Union: viewpoint of the European External Action Service

European External Action Service's (EEAS) Stratcom division was build up to address foreign information manipulation and interference, including disinformation, and contributes to effective and fact-based strategic communication.

« The recent information operation targeting Moldova's presidential election and EU accession referendum has demonstrated how the use of Artificial Intelligence (IA) be functional to FIMI as a whole. In this specific case, AI-generated content and computer program were used to mislead the voters, and influence their perception regarding the EU accession and the presidential candidate Maia Sandu.

A new form of user engagement: the use of AI-enabled chatbot solutions

Right before the event on September 29th, a telegram chatbot was advertised (@NuEuReferend_bot) to mobilize Moldovan citizens to vote "NO" during the upcoming EU Membership referendum. According to the instructions, once the users register to the chatbot, it would assign them "tasks" to complete in order to convince other citizens to vote "NO", and eventually receive a financial remuneration. On September 29th, this chatbot was promoted on Telegram by Ilian Shor (@ilanshor), a Moldovan politician sanctioned by the US and the EU for his actions subverting democracy in the Republic of Moldova (including providing illegal funding to support local pro-Kremlin political activity and "connections to corrupt oligarch and Mosow-based entities"), announced the creation of the chatbot on his telegram account (@ilanshor) on September 29th. Based on EEAS observations, it is the first time in the context of a democratic European election that a chatbot solution offering a direct monetary reward system was used to distribute content and create a call to action aimed at influencing voters' behavior.

Hijacking identity: impersonating individuals with deepfake technologies

Additionally, in early October, an AI-generated video containing a voiceover imitation of President of Moldova, Maia Sandu, was distributed on Telegram and TikTok. This deepfake content alleged that Moldova's accession to the EU would force the country to adopt laws regarding LGBTQ+ community or to sell national land to European foreigners.

The content was then amplified by a Telegram channel affiliated with Moldovan-language version of a Russian media outlet, Komsomolskaya Pravda, and by the website moldova-news.com, which belongs to Portal Kombat, a "structured and coordinated pro-Russian propaganda network", as reported by VIGINUM in 2024.

As highlighted through these incidents, AI's role in information manipulation raise two significant concerns for actors involved in the fight against foreign information manipulation. On one hand, easily accessible AI-powered solutions facilitate content distribution and user engagement through tailored messages targeting specific individuals or groups, thus making the manipulation not only widespread but also personalized. On the other hand, AI technologies like deepfakes and realistic voice synthesis have the potential to ultimately blur the lines between inauthentic and real content and erode citizens' trust among information accessible online.

ii. Sweden: viewpoint of the Psychological Defence Agency (MPF)

The Swedish Psychological Defence Agency (MPF) is tasked with countering foreign information manipulation and interference targeting Sweden and its interests.

Artificial Intelligence (AI), particularly Generative AI, is transforming the landscape of information manipulation, enabling adversarial actors to produce realistic, low-cost, and highly scalable disinformation. This includes fabricated text, images, videos, documents, and sound bites designed to manipulate public perceptions, erode trust, and destabilize democratic institutions. Hostile actors as well as ideologically motivated groups, are increasingly exploiting AI to create tailored disinformation campaigns that resonate with specific audiences, leveraging open-source AI models to bypass resource limitations.

One of the key challenges is mitigating and limiting the spread of AI-generated disinformation, which propagates rapidly and effectively through social media and other digital platforms. This content ranges from sophisticated deepfakes to "cheap fakes," simpler manipulations designed for viral distribution. The increasing accessibility of AI has blurred the lines between authentic and manipulated content, creating what scholars term the "liar's dividend," where even genuine media can be dismissed as fake. These developments complicate fact-checking, increase societal polarization, and pose significant challenges to democratic resilience.

AI is predominantly used for information laundering and the amplification of content. For example, Russian entities have created websites resembling legitimate European and American news outlets. These websites, part of information operations such as the “Doppelgänger” and “Portal Kombat,” serve as “AI-powered portals” controlled by Russian interests. Articles with AI-generated texts hosted on these sites have been disseminated via bot networks and intermediaries on social media platforms, amplified by coordinated networks to maximize reach and impact. These operations often extend beyond social media, aiming to make disinformation campaigns harder to identify and counter.

Ideologically motivated state and non-state actors have leveraged AI to incite division and amplify prejudices. For example, AI-generated imagery depicting caricatured, threatening Muslim figures has been used to fuel Islamophobia, with such content widely disseminated on social media through established hashtags. This strategy aims to normalize prejudice and embed it into mainstream discourse. Similarly, manipulated video clips, such as those portraying American politicians speaking Chinese, may lack sophistication but remain effective tools for spreading disinformation.

Other examples include deceptively edited AI-generated clips and synthetic media productions. Before the U.S. elections, authorities identified the group Storm-1516 as responsible for several high-profile influence campaigns, which relied on fabricated journalists and whistleblowers and manipulated photos. Additionally, synthetic media created by private individuals has exacerbated the oversaturation of the information space, making the detection and countering of disinformation increasingly challenging.

Despite these advancements, the underlying threat remains rooted in traditional influence tactics adapted to new platforms. AI has simply amplified these techniques, making them more pervasive and harder to detect. Countering these challenges requires a well-informed and resilient population capable of withstanding adversarial manipulation. Psychological defence must be a collaborative effort involving government agencies, municipalities, civil society, private organizations, and the general public. This includes raising awareness through targeted campaigns, education, training, and exercises. Research demonstrates that increased knowledge and awareness reduce susceptibility to manipulation, reinforcing the foundations of an open and democratic society.

As the threat landscape grows increasingly complex, with both state and non-state actors leveraging AI to influence international conflicts and democratic processes, the psychological defence must evolve. To enhance its capabilities, MPF has established a Digital Hub to analyze risks, vulnerabilities, and consequences across the digital platform landscape, including social media, AI, gaming, and emerging technologies. This initiative strengthens MPF’s ability to coordinate a cross-sectoral, nationally cohesive defence strategy. Leveraging AI’s defensive capabilities ensures the most effective response, securing the resilience of democratic institutions and societal trust.

iii. Canada: viewpoint of the Rapid Response Mechanism (RRM): Case Study Spamouflage

Attached to Canada's ministry of foreign affairs (Global Affairs Canada), Rapid Response Mechanism (RRM) Canada is responsible for sharing information and analysis on foreign threats to democracies, and identifying opportunities for coordinated response within the G7 RRM.

Canada has watched the threat landscape evolve over the last few years. In particular, the rapid pace of technological advancement, especially artificial intelligence (AI) has further amplified the threats.

Foreign state actors are leveraging both commercial and in-house generative AI in the production of covert and overt foreign information manipulation and interference (FIMI). AI enables these actors to rapidly produce and disseminate synthetic content at scale. The ultimate aim is to sow discord, amplify disagreements and grievances and delegitimize government institutions.

A common tactic exploited by foreign adversarial actors is the use of AI to produce ultra-realistic media, often creating multilingual AI-generated avatars to create and broadcast aligned content in several languages at scale. The experimentation with microtargeting AI-generated presenters, using different accents and skin tones to appeal to local audiences, is highly concerning.

An equally troubling trend is foreign state use of AI to produce deepfake audios and videos to seed false and misleading narratives online, some of which can be difficult to detect and debunk. Of particular concern is the use of these capabilities in the context of gender and identity-based violence, for developing harmful and abusive online content meant to target and discredit individuals, activists, journalists and political figures who are perceived as threats to state actors.

Case Study: Spamouflage

In 2023, Canada investigated a campaign probably related to "Spamouflage", an operation already documented by actors involved in the fight against information manipulation. Several individuals were targeted by this campaign, including political dissidents, Canadian parliamentarians, the Prime Minister and the Leader of the Opposition Party. The campaign's exploited the following tactics:

1. Impersonating individuals though AI-generated videos (deep fake), to accuse dozens of Canadian parliamentarians of criminal and ethical violations.
2. A bot network then leveraged the popularity of verified Canadian parliamentarians' social media accounts.
3. The bot network posted thousands of English and French comments amplifying the deep fake video accusations to the comment section of posts on the verified Canadian accounts.

While the impact of the operation on Canadian parliamentarians was likely low, Canada remains concerned about the use of artificial intelligence in creating and amplifying FIMI to engage in acts of digital transnational repression, and these activities becoming more persuasive and with wider reach.

iv. United Kingdom: viewpoint of the Foreign, Commonwealth & Development Office (FCDO)

Attached to UK's Foreign, Commonwealth & Development Office (FCDO), the cyber information and tech threats directorate aims to provide insights and analysis on the foreign informational threat.

The multi-platform "Newstop" FIMI network, first identified by Meta and evidenced to be using AI by internal OpenAI threat investigators, has been found to be propagating pro-kremlin and anti-Ukrainian narratives to audiences in the UK and francophone Africa.

The network has made use of generative AI tools to create text, image, video, and audio content. This includes videos of AI-generated "news anchors" delivering news stories to audiences on TikTok and AI-generated cartoon images of a political nature included in posts on X.

The UK FCDO has independently found evidence, extending claims made by Meta and OpenAI, that the network operates three distinct clusters of activity. The first cluster operates an Africa-focused brand, targeting audiences in francophone Africa across X, Telegram, TikTok, and YouTube. The Second cluster is a set of English-language assets targeting audiences in the UK across X, TikTok, and news websites, and the third consists of a UK-focused brand targeting UK audiences via a news website with associated X and YouTube accounts.

The network's assets examined by the FCDO have not received significant engagement from users despite being active across multiple platforms.

Using commercially available, closed-source AI services for coordinated operational FIMI activities, access to which is logged and monitored by a third-party for breaches of its terms of service, demonstrates an immaturity of AI capability in this specific instance of FIMI. This practice shows little consideration for operational security and undermines efforts to obfuscate assets used for covert coordinated FIMI activities.

As FIMI actors' capabilities mature, a greater use of open-source AI models running on private computer hardware will reduce the availability of organisational data, such as that from OpenAI, which can be used to triangulate the activities of FIMI actors.

II. Issues and prospects for the impact of IA on the information threat

While information threat actors are increasingly using generative AI technologies to weaponize their manoeuvres, the question remains as to their effects and impact on public opinion.

There is currently no consensus in the academic world on how to measure the impact of a digital information manipulation campaign. Impact analysis is mainly empirical, often consisting of quantitative indicators of visibility, provided by the main social networking platforms (number of views, likes, shares or comments), but these only provide a fragmented view of the exposure of a readership or audience to the campaign, without making it possible to measure its effects over the long term. So, despite the few tools available,[18] little is known about the impact of online information manipulation on the behaviour of target audiences or those exposed to it.

In this section, VIGINUM, based on its observations, sets out to formulate working hypotheses on the risk of impact from malicious manoeuvres using AI.

1. A real risk of a rise in the information threat level, but with a currently moderate impact

a. A risk of increased responsiveness on the part of malicious foreign actors

By reducing the content production time, AI technologies are likely to considerably increase the responsiveness of malicious foreign actors, in addition to enabling them to produce on a large scale. By interacting with a language model such as *ChatGPT*, it is possible to have the tool write quality content that can be mobilised as part of a foreign digital interference operation launched without warning, e.g., for opportunistic reasons. This automation of real-time content creation using a language model requires no special expertise, making it easy to design simple, large-scale operations.

b. A risk of increased stealth of information manoeuvres

AI models that generate text content are now capable of producing texts translated into several foreign languages, with a quality of syntax and usage that sometimes approaches that of a mother tongue. As well as making it easier to reach certain audiences (e.g., diasporas), this performance of GenAI tools can also help to better conceal the involvement of a foreign actor in digital information manipulation campaigns, making it more difficult to detect them.

What is more, when it comes to text generation, tactics that are easy to identify today - such as roughly-translated copy-pasta - could be replaced in the future by more sophisticated processes, in particular by generating a large volume of rewordings, which are more complex to update. In the case of images, the use of image generators to create synthetic content means that new types of advanced detection tools are needed to identify the source.

[18] In particular: *Breakout Scale* by Ben NIMMO (The Breakout Scale: Measuring the impact of influence operations, 2020, https://www.brookings.edu/wp-content/uploads/2020/09/Nimmo_influence_operations_PDF.pdf) and *Impact risk-index* by *EU Disinfolab* (Towards an impact-risk index of disinformation: measuring the virality and engagement of single hoaxes, 2022, https://www.disinfo.eu/wp-content/uploads/2022/06/20220617_IndexImpactAssessment_Final.pdf).

c. A moderate development of the threat at this stage

While GenAI increases the capacity of malicious actors to produce large volumes of content on online platforms, it does not appear at this stage to constitute a real breakthrough in the field of information manipulation.

The quality of the content generated is not necessarily a guarantee of its impact on the target audience. Some research has shown that content manipulated using less sophisticated methods ("*cheapfakes*") can be just as harmful as sophisticated synthetic content.[19] For example, real images, not generated by AI, but instrumentalised or decontextualized as part of information manoeuvres, can have a major impact.[20]

Moreover, while GenAI makes it possible to speed up the production and dissemination of content and lower its cost, it still does not make it possible to resolve the challenges associated with its dissemination and virality among new audiences, which remain the main brake on information manoeuvres.[21] [22]

So, with all due caution, artificial intelligence currently seems to represent more of an "evolution" than a "revolution" in terms of the information threat. It allows existing operating methods to be "industrialised", with much higher production volumes and lower costs, but does not yet seem to have contributed to the creation of new operating methods. Nevertheless, rapid technological developments in AI present structural risks in terms of medium- and long-term information threats.

2. Prospects for the evolution of the information threat linked to AI

a. Towards an altered relationship with reality?

The generalisation and acceleration of GenAI model capabilities raise fears of a massive proliferation on online platforms of content generated by AI tools. This trend can already be seen on certain social networking platforms *(YouTube*, *LinkedIn*, *TikTok*, *Pinterest*) and also concerns the production of synthetic media.

In the long term, this phenomenon could be likely to increase the mistrust of users, and more broadly of the general public, towards all online content, whether authentic or not, and gradually lead to a form of mistrust towards the very notion of information. The rise of these technologies could therefore provoke widespread scepticism (*'deep doubt'*), creating the risk for citizens to profoundly alter their relationship with reality. This is what the authors of a report by the *Institute for Strategic Dialogue* (ISD) on the role of AI in the American elections of 2024 observe: "(...) *very often, the content that people were discussing was not generated by AI, but it was the 'spectre' of AI that had an impact. AI gives people the opportunity to deny reality as they already do, but even more intensely*"[23].

[19] M. HAMELEERS, "Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting", International Journal of Public Opinion Research, 2024, vol. 36, pages 1-9, https://doi.org/10.1093/ijpor/edae004.
[20] For example, in October 2023, 250 stencils featuring Stars of David were found on walls in Paris and the inner suburbs. Photographs of these stencils were then circulated and amplified inauthentically on social networks. The operation was denounced by France: https://www.diplomatie.gouv.fr/fr/dossiers-pays/russie/evenements/evenements-de-l-annee-2023/article/russie-nouvelle-ingerence-numerique-russe-contre-la-france-09-11-23.
[21] A. NARAYANAN and S. KAPOOR, "*The LLaMA is out of the bag. Should we expect a tidal wave of disinformation?*", *AI Snake Oil*, 6 March 2024, https://www.aisnakeoil.com/p/the-llama-is-out-of-the-bag-should.
[22] A. NARAYANAN and S. KAPOOR, "We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem", *AI Snake Oil*, 13 December 2024, https://www.aisnakeoil.com/p/we-looked-at-78-election-deepfakes.
[23] "AI didn't turn the US election upside down, but it did change our relationship with reality", from Radio-Canada. https://ici.radio-canada.ca/nouvelle/2120725/intelligence-artificielle-campagne-electorale-etats-unis-trump-harris

The same ISD report[24] also reveals that users often rely on inappropriate strategies to determine whether or not content is generated by AI. In addition to citizens who are insufficiently equipped, society as a whole is faced with the critical challenge of discerning reality from fiction, and the synthetic from the authentic.

This defiance towards information could be exploited by malicious actors, as researchers at *Yale* University in the United States have theorised as the "*liar's dividend* " syndrome[25]: the more a society learns to be sceptical, the easier it becomes for a liar to question irrefutable facts. This could eventually lead to a large destabilisation of democratic processes.

b. Risks associated with the proliferation of synthetic content

The use of AI and the mass production of content also means that certain malicious actors can occupy online spaces that were previously little used, in order to disseminate their narratives. The ability to produce content on a large scale means that large-scale campaigns can be rolled out, positioned for the long term, and targeted at specific audiences. We are therefore likely to see a proliferation of networks of fake local media, like *Operation PAPERWALL*, whose operating method consists of creating sites imitating local news portals in order to disseminate propaganda and disinformation to local targets.[26] AI could also be used by malicious actors to promote the search engine optimisation[27] of certain content on very specific subjects.

Furthermore, on platforms such as *YouTube* or *TikTok*, where the recommendation algorithm is particularly powerful and offers users content similar to that with which they have already interacted, the proliferation of similar content on a given topic is likely to accelerate the 'rabbit hole' phenomenon by trapping users in a bubble of potentially manipulated content.

c. Risk of pollution and self-degradation of AI models

The proliferation of the use of GenAI services also raises the question of the biases contained in the training data used, as well as that of the sources potentially mobilised to generate the response. When it comes to training models, suppliers of pre-trained models are often opaque about the data used and the potential bias in the responses of the GenAI. The training data is the result of collection, filtering, and processing choices, which are reflected in the model generations. Part of the scientific community is working to make such models more transparent,[28] and is carrying out work to measure biases in the data sets used,[29] as well as in the responses of GenAI models.[30]

To avoid providing misleading responses, some services based on GenAI models suggest sources in support of their response, but without first analysing the quality of the information they provide. For example,

---

[24] *Institute for Strategic Dialogue (ISD)* report "Disconnected from reality: American voters grapple with AI and flawed OSINT strategies", dated 7 November 2024. https://www.isdglobal.org/digital_dispatches/disconnected-from-reality-american-voters-grapple-with-ai-and-flawed-osint-strategies/

[25] Kaylyn JACKSON SCHIFF, Daniel SCHIFF, and Natalia S. BUENO, 2024, "*The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?*", *American Political Science Review*, https://doi.org/10.1017/S0003055423001454

[26] https://citizenlab.ca/2024/02/paperwall-chinese-websites-posing-as-local-news-outlets-with-pro-beijing-content/

[27] Search engine optimisation (SEO) is a strategy for improving the visibility of a website on search engines in order to reach a target audience.

[28] *The Foundation Model Transparency Index* from Stanford University https://crfm.stanford.edu/fmti/May-2024/index.html

[29] For example, the *Knowing Machines* project is analysing the construction biases of the LAION 5B dataset, used to train certain image models: https://knowingmachines.org/models-all-the-way

[30] For example, an analysis shows that the responses provided by the Qwen 2 7B artificial intelligence model developed by *AliBaba* does not reflect the universally shared views on certain political issues https://huggingface.co/blog/leonardlin/chinese-llm-censorship-analysis

Newsguard[31] has shown that certain GenAI services use sites from the pro-Russian *Portal Kombat* system detected and characterised by VIGINUM[32] as a source for generating their responses.

In addition to the threat of deliberate attacks on data integrity, the proliferation of generated content online is likely to pollute the training data that is retrieved in large-scale *web* collection operations. According to some researchers, the proliferation of such content could lead to a sharp deterioration in the performance of AI models, as it becomes increasingly integrated with their training data.[33]

[31] https://www.newsguardtech.com/ai-monitor/french-language-ai-misinformation-monitor/
[32] 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
[33] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal, 2024, "*AI models collapse when trained on recursively generated data*", *Nature,* https://www.nature.com/articles/s41586-024-07566-y

III. AI as an opportunity to support operations to fight information manipulation

While artificial intelligence can be used maliciously as part of information manipulation campaigns, it can also be used for more noble causes, and in particular it can be extremely effective in analysing inauthentic behaviour and other operating method.

In this respect, recent developments in artificial intelligence, and in particular the availability of pre-trained models, offer those involved in the fight against information manipulation new possibilities for exploring, analysing, and characterising profiles on online platforms, text, audio, visual, or video content, or the dynamics of disseminating a narrative on social networks.

1. The experience of VIGINUM

a. VIGINUM's Datalab: advanced data analysis to support operations

To strengthen its investigative and analytical capabilities, VIGINUM has invested resources in technological innovation, by setting up a Datalab as soon as it was created in 2021. In addition to its primary mission of supporting the department's investigations, VIGINUM's Datalab carries out research and development (R&D), enabling it to implement innovative methodologies to support operational missions, as well as publishing academic articles and, for the first time in 2025, open-source software. The Datalab's R&D activity not only provides technical and analytical support for the department's investigations, but it also equips the entire community involved in the fight against information manipulation (civil society, researchers, media, etc.).

As part of this R&D work, the Datalab implements a wide range of data analysis methods and tools, including solutions using AI. While AI can make a significant contribution to analysis (see below), it is not a miracle solution, and its use requires solid technical expertise, combined with a good understanding of the strengths and limitations of each model. The Datalab therefore adopts a practical, pragmatic approach to AI, which represents a set of models and tools that can be mobilised according to need, as part of a wider range of solutions.

Many of the data processing operations carried out by the Datalab do not fall within the scope of AI. For example, the traffic manipulation coefficient proposed by Ben NIMMO is a simple indicator that measures the amplification of a hashtag on the *X* platform, and is not based on AI.[34] Similarly, the methodology developed by VIGINUM to detect trending topics on the *X* platform is based on simple statistical anomaly detection methods.[35]

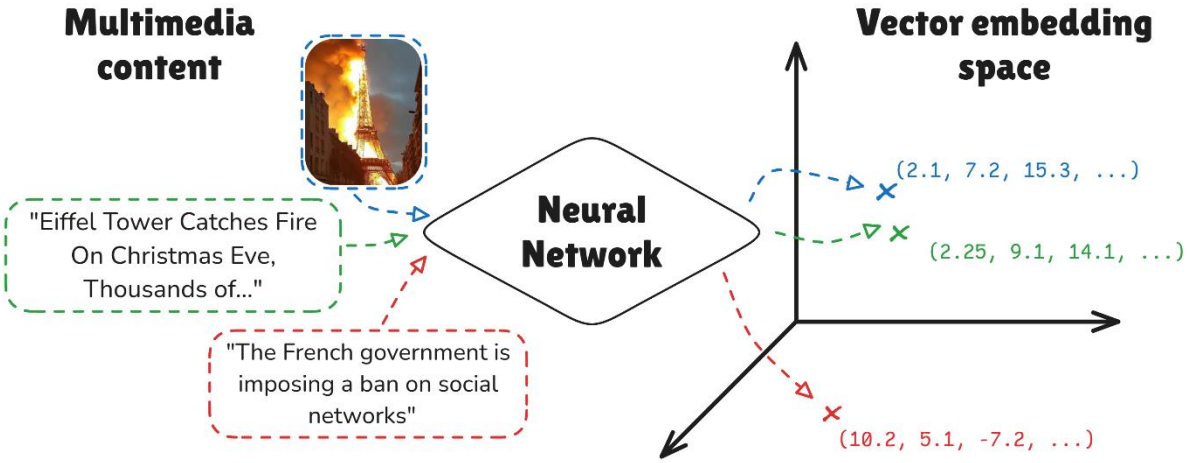
[34] Ben NIMMO, "Measuring Traffic Manipulation on Twitter." Working Paper 2019.1. Oxford, UK: Project on Computational Propaganda. 35 pp., https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2019/01/Manipulating-Twitter-Traffic.pdf

[35] Cristian BROKATE, Manon RICHARD, Lisa GIORDANI, and Jean LIÉNARD. 2024. "SWATTING Spambots: Real-time Detection of Malicious Bots on X", Companion Proceedings of the ACM Web Conference 2024 (WWW '24), Association for Computing Machinery, New York, NY, USA, 818-821. https://doi.org/10.1145/3589335.3651564

Examples of the use of AI implemented by VIGINUM's Datalab:

***Direct use of a model through a* prompt**

As generative models (*Mistral 7B instruct*, *Llama*, *Gemma*) are pre-trained on a massive corpus of texts, they contain a body of knowledge on a large number of specialised domains. This gives them good zero-shot learning capabilities,[36] which means they can be used directly via prompts. One of the problems encountered with this type of use is the variability of the model's responses, as well as their formats, depending on the input prompt.[37] As part of its operations, Datalab uses prompt engineering, e.g., to detect the countries mentioned in posts on social networks. This is particularly useful for filtering posts specifically targeting France within a given corpus.



A vector embedding is a representation of the semantics of content in a mathematical space.

Using vector embeddings

A vector embedding is a digital representation of a document (text, image, audio, etc.) that captures its semantics. AI models trained on massive corpora of documents can thus represent heterogeneous documents as a fixed list of numbers in a mathematical space.[38] This representation can then be used to apply other algorithms, such as calculating distances between documents, detecting semantically similar images or texts, or segmentation. It is used by the Datalab in its methodology for detecting duplicated content.[39]

Model fine tuning

Model fine tuning is used to specialise a pre-trained AI model on a task and an annotated dataset. This makes it possible to use a model's generalist knowledge to specialise in a given task. For example, the Datalab has used this principle to automatically classify ads in the register of *Meta*[40] ad content.

[36] Zero-shot learning is the ability of an AI model to make predictions about tasks or concepts without ever having seen them during its training.
[37] The responses of a generative model can sometimes respect the response structure imposed by a prompt without respecting the instructions.
[38] For example, the well-known BERT model transforms a sentence into a vector of 768 numbers.
[39] Richard et al, "Unmasking information manipulation: A quantitative approach to detecting Copy-pasta, Rewording, and Translation on Social Media", 2023, https://arxiv.org/abs/2312.17338
[40] See the work presented at FOSDEM 2024: https://archive.fosdem.org/2024/events/attachments/fosdem-2024-3204-detecting-propaganda-on-facebook-and-instagram-ads-using-meta-api/slides/22323/Fbads_FOSDEM_20240203103844_MlWExhL.pdf

b. Using AI to combat foreign digital interference

The Datalab uses AI both to facilitate data mining and to identify markers of inauthenticity, in order to characterise foreign digital interference.

i. Semantic analysis of text content

As part of its mission to detect and characterise foreign digital interference, VIGINUM analyses text data in natural language, such as messages collected on social networking platforms *(X*, *Facebook*, *Telegram*, *Threads*, etc.*)* or websites (blogs, fake media, etc.). In this field, the progress made over the last ten years in automatic language processing, and in particular the development of pre-trained language models, has made it possible to achieve significant gains in many automatic natural language processing tasks, particularly with regard to the semantic analysis of text content.[41]

For example, as part of its support for operations, the Datalab regularly uses topic modelling models, which are based on vector embedding models, as well as models for recognizing named entities.[42]

Topic modelling models such as BERTopic can be used to detect the topics covered in a corpus of texts, and to group together content dealing with the same topic.[43] These methods make it possible to identify the topics addressed by a threat actor ('narratives') or a set of actors in a corpus, to determine whether the subjects addressed are of interest to the service, or to target a subset of interest. As part of our investigative support, the detection of topics is an important part of the exploratory analysis of textual content. The Datalab's ongoing technology watch, particularly on the subject of topic modelling, enables it to keep abreast of the latest developments.

In addition to detecting topics, automatic language processing can also identify named entities within a corpus of textual content. For example, the Datalab has set up a processing chain to detect messages that refer to a given country, either by mentioning it in full or by mentioning one or more of its localities. This processing chain makes it possible to identify the publication of content concerning France by foreign information threat actors. Among the methods tested, some are based on reconciling entities with an existing knowledge base, in order to link localities to a given country.[44] Other methods rely on the instructions (prompts) given to large language models (such as *Mistral 7B*) to identify the countries associated with localities mentioned in a text.

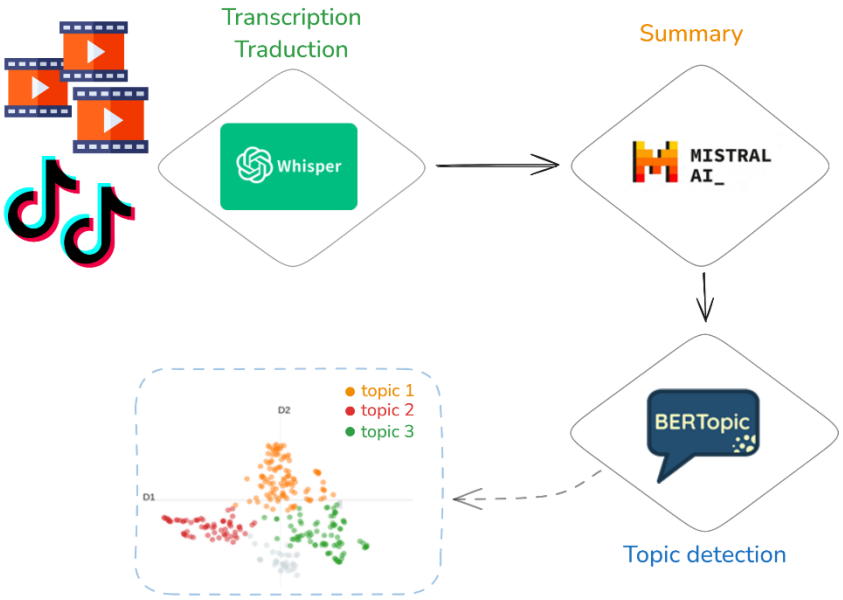
[41] Published in 2013, the Word2vec model has met with great success by making it possible to project words into a semantic space (Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013, "Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 [cs.CL]). From 2018, pre-trained language models such as BERT (Kenton, J. D. M. W. C., & Toutanova, L. K., 2019, "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of naacL-HLT*, vol. 1, no. 2.) and then GPT can be used to model the meaning of words according to their context. For a history of language models up to *ChatGPT*, see Pierre-Carl LANGLAIS (February 7, 2023), "*ChatGPT*: how does it work?" *Sciences communes*, https://doi.org/10.58079/twxr

[42] Named entity recognition is an information extraction task that seeks to locate and classify elements in text into predefined categories such as a place, a person, an organisation, a date, etc.

[43] https://github.com/MaartenGr/BERTopic

[44] For example, the Spacyfishing method (https://spacy.io/universe/project/spacyfishing) can be used to associate an entity with the *Wikidata* knowledge base.

ii. Building a processing chain to explore video content



Processing chain combining different artificial intelligence models to explore the content of a corpus of videos.

The exploratory analysis of video content, e.g., from the *TikTok* or *YouTube* platforms, is generally more complex for analysts than text content analysis. Nevertheless, it is possible to use large video and language models[45] or large multimodal models to analyse video content, but it is also possible to use simpler methods, by focusing on analysing the video soundtrack.

By combining different AI models, the Datalab has, for example, set up a data processing chain to support an investigation into the exploratory analysis of a set of *YouTube* or *TikTok* channels, using only soundtrack analysis. This processing chain combines a transcription unit, a translation unit, an automatic summarisation unit, and a topic detection unit. For transcription, open source templates such as the *WhisperAI* template family[46] can be used. This unit can then be combined with a large language model to obtain an automatic transcription summary for each video.

Since 2023, many open-weight language models, such as *LLama*,[47] or open source models such as *Mistral7B*[48] or *Mixtral8x7B*,[49] can be used to summarize the content of transcriptions. Topic modelling models (see above) can then be used to group together videos dealing with the same topics. Combining these pre-trained models facilitates the exploratory analysis of a corpus of videos.

[45] For example, https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
[46] https://openai.com/index/whisper/
[47] https://github.com/meta-llama/llama3
[48] https://huggingface.co/mistralai/Mistral-7B-v0.1
[49] https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

The importance of open source models and controlling model hosting

Some of the major pre-trained AI models are closed, and only made available to users as a service by the company that developed them. Others are publicly accessible under an open source licence, such as Mistral 7B, made available by the *Mistral* company, or a more restrictive licence, such as the *LLama* models made available by *Meta*, whose weights are open but whose uses are restricted by the specific licence.[50]

For services involved in fighting information manipulation, it is important to control the conditions under which their tools are hosted, for reasons of IT security, operational security, personal data protection, and sovereignty issues. VIGINUM has its own IT infrastructure to host data and carry out processing in a controlled IT environment, on servers located in national territory.

If private providers of AI services can implement safety measures against malicious uses, it is more complex to secure the use of downloadable models, such as open source models. Nonetheless, the development of open source models or models with open weights is a major challenge to enable the actors and entities involved in the fight against information manipulation to use pre-trained models in a controlled IT environment.

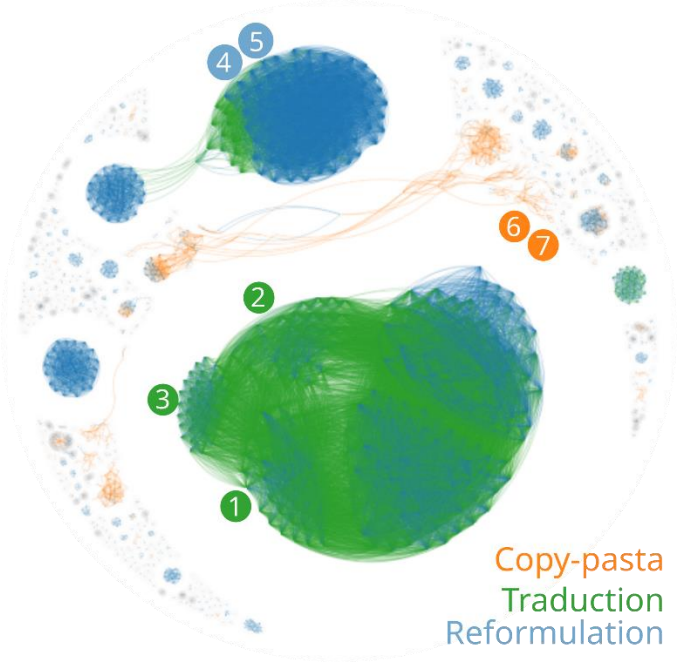
In addition to exploratory content analysis, AI tools are also used to identify markers of inauthenticity, e.g., by identifying massively duplicated content.

[50] As PEReN points out, the definition of open source in the field of AI is not yet consensual. The *Open Source Institute (OSI)* proposed a definition in October 2024, but few models comply with the *OSI* definition (https://www.peren.gouv.fr/actualites/2024-10-29_comparateur_iag_open_source/).

iii. Detecting massive content duplication

Analysis of a corpus of messages using the 3-delta method.

VIGINUM applied the 3-delta method to the "Venezuela 2021" corpus provided by the Twitter Transparency Center[51] and identified groups of similar messages.



Traductions :

- 1 : « ¿Cuándo cumplirá Cabo Verde con la orden de Ecowas [...] »
- 2 : « when will Cabo verde comply with Ecowas order [...] »
- 3 : « Cabo verde acatará a ordem do Ecowas [...] »

Reformulations:

- 4 : « #TWDxSTARChannel ya falta muy poco tiempo para el gran momento »
- 5 : « #TWDxSTARChannel ya por favor que el tiempo pase mas rapido solo por hoy #TWDxSTARChannel »

Copy-pasta:

- 6 : « Hoy toda Venezuela es Alex Saab exigimos liberen [...] de nuestro embajador Alex saab free Alex saab https://t[.]co/M1IJn8JEwg alex Saab »
- 7 : « Hoy toda Venezuela es Alex Saab exigimos liberen [...] de nuestro embajador Alex saab https://t[.]co/msEQgzrfFK alex Saab »

To disseminate a textual message, information threat actors can use a variety of methods: massive duplication or "copy-pasta"; rewording; or translation. To detect these three operating modes, the Datalab developed the 3-delta methodology.[52] Using a pre-trained language model, *Universal Sentence Encoder*,[53]

[51] In 2021, *Twitter* (now *X*) identified and deleted a network of 277 accounts that massively disseminated narratives favourable to the Venezuelan government. The dataset was published by the *Twitter Transparency Center*. See https://fsi.stanford.edu/news/twitter-takedown-december-2021

[52] Richard et al, "Unmasking information manipulation: A quantitative approach to detecting Copy-pasta, Rewording, and Translation on Social Media", 2023, https://arxiv.org/abs/2312.17338

[53] YANG, Y., CER, D., AHMAD, A., GUO, M., LAW, J., CONSTANT, N., ... & KURZWEIL, R. (2019). "*Multilingual universal sentence encoder for semantic retrieval*". arXiv preprint arXiv:1907.04307.

the 3-delta methodology detects pairs of messages with a high degree of semantic proximity. The method is thus able to identify pairs with high graphical proximity and high semantic proximity, considered copy-pasta, pairs with high semantic proximity, the same language, and lower graphical proximity, considered rewording, and pairs with high semantic proximity, lower graphical proximity, and a different language, considered translation. This method detects groups of messages that are abnormally similar, and may therefore have been duplicated in an inauthentic manner.

As part of the AI Action Summit, organised by France on 10 and 11 February 2025, VIGINUM is publishing the D3lta software library[54] for the first time under an open licence, to enable the various players in the ecosystem involved in the fight against information manipulation to reuse and improve this method.

iv. Detecting bots

The undeclared use of automated accounts (or bots[55]) is a marker of inauthenticity. In addition to relatively simple detection heuristics, relevant for bots with regular behaviour (post times and frequency, similar posts, etc.), AI can be used to classify other types of account as bots. The Datalab has built its own bot classifier on *X*, based on an analysis of the network of followers and following on a given account.[56]

v. Detecting AI-generated content

Faced with the growing use of GenAI by information threat actors to create content, those involved in the fight against information manipulation need to equip themselves with tools that enable them to detect synthetic content, either to be able to distinguish synthetic content from authentic content, or to characterise the operating method used. Like GenAI tools, synthetic content detectors are themselves tools that use AI methods. For example, the *Binocular* method, which detects content generated by large language models, itself relies on large language models to measure the probability that a text has been written by a human or an AI.[57]

There are numerous AI content detectors, whose performance generally varies depending on the type of content analysed. This often means users testing multiple detectors to determine the synthetic nature of a given piece of content. With a view to simplifying the process of detecting AI texts and images for the general public, VIGINUM and PEReN have joined forces to invite researchers and experts to participate in the development of a digital commons, consisting of an open-source meta-detector for artificial content, built on the basis of realistic examples observed during information manipulation campaigns.[58]

[54] https://github.com/VIGINUM-FR/D3lta
[55] Bot: automated computer programme designed to simulate human behaviour on social networks. A bot is capable of posting, commenting, sharing, or liking other posts.
[56] The approach is based on a convolutional neural network method on heterogeneous graphs developed by Feng et al. 2021, "BotRGCN: Twitter Bot Detection with Relational Graph Convolutional Networks", arXiv:2106.13092 [cs.SI].
[57] HANS, A., SCHWARZSCHILD, A., CHEREPANOVA, V., KAZEMI, H., SAHA, A., GOLDBLUM, M., ... & GOLDSTEIN, T. (2024). Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. arXiv preprint arXiv:2401.12070.
[58] https://code.peren.gouv.fr/open-source/ai-action-summit

2. Experiences from civil society

As a complement to VIGINUM's perspective, this section details AI use cases implemented by civil society actors involved in the fight against information manipulation.

a. Viewpoint of Lupa, a Brazilian fact-checking agency

The *Lupa Newsroom*, a Brazilian organisation that fights disinformation and carries out fact-checking activities, will be ten years old in 2025. Over the years, *Lupa* has witnessed the growing sophistication of disinformation, both in its strategies and in its speed and scale. These changes have created new challenges for fact-checkers and NGOs involved in the fight against information manipulation, who must now provide verified information to the public faster than misinformation spreads - because it is during this time that 'fake news' causes the most damage.

But it's not just a question of speed. Faced with the growing sophistication of the information threat, tools must also be strengthened to improve fact-checking capabilities, in particular to analyse, extract data, and produce in-depth content analyses in an ocean increasingly polluted by misleading information.

With this in mind, *Lupa* has been developing its own public discourse monitoring tool since 2023. Thanks to *LupaScan*, journalists and researchers can analyse the statements made by 596 elected Brazilian federal officials (president, vice president, and federal senators and deputies). In 2024, the tool became even more robust, incorporating several functions based on artificial intelligence (AI). Users now have access to a dashboard that automatically displays graphics and other visual data, providing information for reports and research. One of the new features includes AI-assisted facial recognition, and the grouping of photos posted by political figures, making it possible to analyse a large set of posts containing images, even when the name of the person depicted is not mentioned in the message.

One of the most remarkable features is the real-time transcription tool. This tool has enabled *Lupa* to become more agile and reduce the risk of errors when covering events, such as debates and interviews, as manual transcription is no longer necessary. The transcription tool works in Portuguese and English, and will soon be extended to other languages.

This tool will not be limited to the *Lupa* newsroom. The organisation plans to establish partnerships, particularly with Latin American organisations, as several countries in the region will be holding elections in 2025.

b. Viewpoint of Full Fact, a UK independent fact-checking organisation

« **Developed by independent UK fact-checking organisation *Full Fact*, *Full Fact AI* provides fact-checking organisations with tools to combat misinformation on a massive scale.** Deployed in 45 organisations across 26 countries, and available in English, Arabic, and French, these tools help fact-checkers monitor public discourse, identify false claims, and combat repeated disinformation. The tools cover a range of platforms, including online news sites, RSS feeds, parliamentary archives, social networks, radio, television, podcasts, and *YouTube*. Each day, the content is converted into text, broken down into sentences, and enriched with additional information. This process helps fact-checkers to manage vast quantities of data on a scale exceeding human capacity.

Daily data is compared with previous checks to detect repeated misleading information, even if it is worded differently. Fact-checkers can then quickly adapt existing work and provide accurate information.

One of *Full Fact AI* 's most recent and most innovative solutions is its generative health misinformation detection tool. This tool monitors multimodal content on all platforms, by analysing captions, images, sounds, and on-screen text. By ranking false health information in order of importance, the tool enables organisations to prioritise resources and respond quickly to dangerous allegations, even when the false information is implicit rather than explicit.

Future developments include automating the monitoring of channels prone to misinformation, and providing daily, keyword-based snapshots of harmful disinformation. These features will improve fact-checkers' workflows, allowing them to concentrate on combating the most dangerous misleading information.

Such tools are no substitute for fact-checkers. By combining cutting-edge AI with human expertise, *Full Fact* can increase the reach and impact of fact-checkers, without compromising accuracy. *Full Fact AI* demonstrates how technology can streamline workflows, improve accuracy, and boost confidence in societies."

The screenshot shows Full Fact *'s* GenAI *prototype, which scans online videos for misleading health information, and provides a list of paraphrased claims for quick and easy review by the organisation's fact-checkers*

Conclusion

While AI offers opportunities in many areas, it also represents a challenge in the fight against information manipulation, and poses a continuum of risks for the digital information space.

On the one hand, although the examples documented in open source tend to demonstrate that the use of AI does not, for the moment, facilitate the propagation of an information manipulation campaign or increase its impact, the growing use of this technology could lead to a structural increase in the information threat, in that it increases the responsiveness of malicious actors, as well as the scale and stealth of their actions.

On the other hand, AI's ability to generate credible fake content poses the risk of widespread public scepticism about the authenticity of any type of online content. If authenticity were to become more easily disputable, our relationship with reality could be profoundly altered. At a time when the general public's trust in information seems to be weakening, this phenomenon is likely to further increase the mistrust of certain citizens towards the media and fact-checkers, creating a breeding ground for those who manipulate information.

In the face of these unprecedented risks, AI itself represents an effective solution for strengthening the fight against information manipulation, whether for exploring massive, varied data, or for detecting inauthentic behaviour. To increase our collective defensive capabilities, it is essential that all the players involved in this fight are able to innovate, cooperate, and share high-performance analytical tools.

This is the approach taken by VIGINUM, which has decided to publish for the first time - as part of the AI Action Summit on 10 and 11 February 2025 - the D3lta software library under an open licence, to enable the various stakeholders to use and improve this method.

ABOUT VIGINUM

Created on 13 July 2021 and attached to the SGDSN (General Secretariat for Defence and National Security), VIGINUM is tasked with protecting France and its interests against foreign digital interference.

The role of this national technical and operational service is to detect and characterise information manipulation that involve foreign actors and aims at harming France and its fundamental interests

Service de vigilance et protection contre les ingérences numériques étrangères | SGDSN

Cover photo credit: *photo and machines* on Unsplash