DOI: 10.53477/1842-9904-24-21

LARGE LANGUAGE MODELS, PROPAGANDA AND SECURITY CHALLENGES

Diana-Cristiana LUPU, PhD*
Daniela LICĂ, PhD**

The present paper is a non-systematic narrative review of security challenges and solutions related to the LLM-generated propaganda, considered in the context of influence activities. The purpose of the paper is to provide a synthesis of the knowledge on the mentioned topic, based on research, opinion and regulatory documents published between 2017 and 2024. To that end, the developed research protocol is designed to take into account criteria related to the diversity, credibility and eligibility of primary and secondary sources. The synthesis of topic-related knowledge is then illustrated and discussed in a manner as objective as possible. Thus, we consider that the main findings can be of help for researchers to identify, justify and refine hypotheses, focusing on possible pitfalls and gaps, as well as for the general public to acquire a higher level of situational awareness, given the novelty of the topic. Moreover, they may contribute towards targeting new avenues for research in the field.

Keywords: *generative AI; large language models (LLMs); propaganda; influence activities; AI-related security challenges.*

Introduction[1]

Large Language Models (LLMs) are algorithms that, based on very large data sets or big data, recognize, summarize, translate, predict, and generate content.

[1] A.N.: A synthesis of this research was presented at the International Conference *Societal and Technological Determinants of Security: Adaptive Strategies for Information and Cyber Challenges*, organised on 24 May 2024 by War Studies University, Poland, and İzmir Kâtip Çelebi University, Turkey (online participation).

**** Diana-Cristiana LUPU, PhD., works for the Romanian Military Thinking Journal, within the Romanian Defence Staff. E-mail: diana_c_lupu@yahoo.com***
***** Daniela LICĂ, PhD, is Researcher at the Centre for Defense and Security Strategic Studies within "Carol I" National Defence University, Bucharest, Romania. E-mail: lica.daniela@unap.ro***

They represent a class of deep learning architectures that are called transformer networks, the transformer model referring to a neural network that learns context and meaning by tracking relationships in sequential data, which can be the words that form a sentence. Following the model of natural language, a transformer is made up of multiple transformer blocks, also known as layers, thus being able to decipher input and predict output, in relation to feed-back and normalization processes (Nvidia, 2024).

Transformer architecture was first discussed by Google in 2017, when eight researchers published a study focusing on attention mechanisms and expanding them into machine deep learning processes. In the 7-year period since its release to the public, the study, dubbed *Attention Is All You Need*, has been cited for over 100,000 times, being considered the modern artificial intelligence founding document. The researchers present the Transformer as the first model completely based on attention, replacing the recurrent layers most frequently used "in encoder-decoder architectures with multi-headed self-attention", being thus faster and more effective than other architectures based on recurrent or convolutional layers (Vaswani & al, 2017, p. 9).

Considering the complexity of human communication science, the introduction of another actor, which is non-human, although it is trained by humans, has complicated the already inter- and trans-disciplinary research paradigm, providing the focus on competition for information and attention with a completely new perspective. The possibility of automated content creation has also resulted in reconsidering influence activities, generating extensive discussions on the recently introduced vulnerabilities, as well as on the necessity of adopting new rules and regulations in the field.

Against this backdrop, the present paper, employing empirical and qualitative research methods, is intended to be a non-systematic narrative review of the identified security challenges and solutions related to the LLM-generated propaganda. Propaganda is considered in the context of influence activities intended to alter beliefs, generate social polarization, change views and voting behaviour or inspire political and other types of violence.

Thus, the purpose of the article is to provide a synthesis of the knowledge on the mentioned topic, based on the research, opinion papers and regulatory documents published between 2017 and 2024. We consider it valuable for the following reasons. Firstly, the topic is a relatively novel one and a synthesis of the main findings can help the general public to acquire a higher level of situational awareness. Secondly, a literature review can help academia to find new directions of research, refine hypotheses, and identify possible drawbacks and gaps in the literature. Thirdly, it could help decision-makers to consider broader and interconnected perspectives in order to adopt well-documented and responsible measures. Having in view the objectivity criteria, we have developed a research protocol intended to mitigate the risk of bias, which is detailed as follows.

Methods

The preliminary search of the literature has helped us to refine the topic and objective, as well as to establish the most appropriate research protocol, having considered the need for structuring the information presentation to meet criteria related not only to content diversity, objectivity and credibility, but also to editorial requirements. Thus, the protocol has been focused on finding the sources of information, setting the specific parameters for the literature search, as well as on establishing the selection criteria, namely the inclusion, exclusion and compilation ones.

With regard to the sources of information and the specific parameters for the eligible literature, we have searched the internet using keywords such as LLMs, challenge, security, propaganda, influence activities, regulation. In this way, we have created a bibliography list to be studied, focusing on the review topic, objectives and possible future developments.

Considering the great number of results, we have established some selection criteria, as follows. In terms of the period covered by the research, we have decided it to be that between 2017 and May 2024. The beginning of the period coincides with the publication of the study *Attention Is All You Need* (Vaswani & al, 2017) considered, as it has been previously shown, the modern artificial intelligence founding document. Thus, the research does not cover the already-documented studies related to the 2016 US elections.

Regarding the perspective diversity in relation to guarding against bias and credibility, we have selected information provided by LLM developers, academia, think tanks and international organizations such as the EU, focusing on the models present and prospective performance, the identified challenges and the already in force regulatory framework in the field. Thus, following the identification of pertinent information and its coding, considering the amount of data, we have applied further selection criteria, mainly intended to restrict the area of research to the most relevant examples, to streamline redundant information, and to provide an as objective as possible synthesis, readable by both experts and general public, as we consider the topic should be of interest for everyone who wants to make responsible decisions in an age increasingly and profoundly affected by the AI flying progress.

Discussion

Considering the relative novelty of the topic and the varied degree of the audience familiarity with it, as well as the fact that there is still limited consensus on the definition of propaganda (Licǎ, 2024) in the context of its global dynamics, we have provided a synthesis of the most relevant identified information, developed in compliance with the described research protocol. Thus, the LLMs functioning principle and their potential to generate propaganda, seen in the context of influence activities, are explained. The major areas of concern in relation to security, be

they challenges, vulnerabilities or even threats, are revealed and discussed. The potential solutions to the identified issues are then presented, showing that they can be provided by means of the same mechanisms that generate the problem, by regulatory frameworks, or by an intelligently-agreed mix of them. As for already existing regulatory frameworks in the field, the EU example is provided. All the above-mentioned aspects are accompanied by the authors' logical interpretation, taking into account limitations related to comprehensiveness, systematization, and possible biases of the present narrative review.

1. LLMs and the Potential to Generate Propaganda

LLMs are systems that employ deep learning techniques, based on very large data sets or big data, to generate content, following the model of natural language processing (NLP). Although studied since the '60s, they became more prominent in 2017, when the neural networks called transformers were developed, allowing for their great efficiency and effectiveness and thus for their applicability in many domains. In 2021, the Stanford Institute for Human-Centered Artificial Intelligence (HAI) coined the term *foundation models*, referring to models trained on broad data, which can be adapted to a wide range of tasks, being thus the subject of a paradigm shift. They are homogenization intensifiers, based on consolidation. However, consolidation is considered a double-edged sword, as it can reduce bias and improve robustness, on the one hand, and it identifies these models as singular points of failure that can radiate harms such as security risks and inequities to countless downstream applications, on the other hand (Bommasani & Liang, 2021). The foundation models need, therefore, to be trained on a large volume of data. Moreover, as they follow NLP, we can understand that these models, LLMs included, are both pre-trained and trained in use, by means of feedback. Since 2017, many such models have been developed, the most popular ones being ChatGPT, developed by OpenAI, and Gemini, developed by Google, some of them being free to use.

It is thus evident, considering the ability of such models to create content, that there is the possibility for the particular content to be harmful. However, within the context of this paper, we will focus on the propaganda-related harm, propaganda being used here in the context of influence activities, as communication intended to further someone's malicious agenda, by determining the message receiver to choose a certain response that advantages the sender. It is also known as reflexive control. In NATO terms, propaganda refers to "information, ideas, doctrines, or special appeals disseminated to influence the opinion, emotions, attitudes, or behaviour of any specified group in order to benefit the sponsor, either directly or indirectly" (NATO, 2020, p. 258), while the hostile content is that "produced by an adversarial actor, either openly or covertly, with the express purpose of countering NATO's message

and mission" (NATO, 2020, p. 152), disinformation and propaganda, as well as media trolls and bots being provided as examples. The main issues related to LLM-generated propaganda consist in the fact that they are targeted to the receiver, based on the analysis of existing data related to the receiver's habits and preferences, on the one hand, and in the speed at which the message is propagated and even amplified, on the other hand.

In this context, the inherent mechanisms involved in LLM-generated content should be also considered as potential risk factors in relation to disinformation and propaganda, namely *pre-training, fine-tuning* and *reinforcement learning* from human feedback (RLHF) and reinforcement learning from AI feedback (RLAIF). As far as *pre-training* is concerned, mention should be made that models can be tricked, then propagating the trick. Moreover, LLMs learn to predict the next word by recognizing patterns in enormous quantities of text data, a process called *self-supervised pre-training*. In addition, the process also entails *representation* and *transfer learning*, referring to the situation in which a model trained for one task seems to be able to *transfer* what it has learned to a different task, based on a useful *representation* of text from its training data. Thus, such a pre-trained model will produce text that mirrors the internet text it has been trained on, which can be malicious in itself, such as hate speech, dangerous information etc., a problem that AI developers have to solve (Burtell & Tonner, 2024).

With regard to *fine-tuning*, it is one of the solutions provided by developers to limit the production of harmful outputs, either by mimicking malicious characteristics of data or by producing plausible yet false outputs. Thus, *fine-tuning* refers to methods of refining pre-trained models to meet particular purposes, entailing tasks or instructions, while *reinforcement learning* is an approach focused on training models to maximize a chosen measure of task performance, referred to as a *reward model*, namely an indicator of how effective the LLMs performance is to meet the task requirements. To do this, LLM developers collect data specifically for training their reward models, using RLHF, which entails human annotators to compare, rank and choose different responses, thus incorporating bias, or RLAIF, which resembles RLHF, the difference being that the LLM itself is provided with written instructions on how to rate outputs. In this context, it is hope that, in the future, LLMs may be able to evaluate each other on complex tasks that exceed the human capacity of unbiased or complete and complex evaluation. However, these techniques provide no guarantees of model behaviour, especially in untested situations, not being based on principled situations. In addition, they can be circumvented by malicious actors, seeking to either "trick" models into providing harmful responses or "undo" the initial fine-tuning, especially considering that LLMs developers, although they have some tools to control the output, such as filters, in order to meet specifications, are not flawless and unbeatable (Woodside & Toner, 2024).

To the above-mentioned aspects related to inherent LLMs issues, the one of LLMs hallucinations can be added. It is a phenomenon in which a "LLM, chatbot or computer vision tool perceives patterns or objects that do not exist or that are imperceptible by people, creating outputs that are nonsensical or altogether inaccurate" (IBM, 2024). It can occur as AI algorithms sometimes "produce outputs that are not based on training data, are incorrectly decoded by the transformer or do not follow any identifiable pattern", because of training data inaccuracy or bias and high model complexity (IBM, 2024). In legal context, hallucinations of models are incriminated in up to 75% of the cases (Dahl, 2024). As for the relationship between LLMs hallucinations and propaganda, the inaccurate, false or disconnected content, generated, even unintentionally, because of the internal vulnerabilities of the model, may be employed, following its multiplication, in pursuing discriminatory, polarized or violent agendas, considering the possibility of altering the decision-making process. In this context, there can be unforeseen and undesirable consequences of using AI tools, which acquire more importance, considering especially the open-source technology, case in which preventing such issues has become increasingly challenging. Although developers claim to have addressed and resolved most of the identified hallucination-related issues, it is obvious that there may be others not yet identified, emerging, or, in the context of the present paper, intentionally invoked by malicious users, which raises the question of accountability.

Another largely debated topic-related issue is that of LLMs already existing or possible to exist autonomy. Although the number of experts who claim that AI autonomy belongs to a distant future, the prospect should be carefully considered, as the line between promise and peril could be extremely thin and blurred. On the other hand, it is natural for developers to want more and, according to available data, OpenAI, Microsoft and Google autonomy-related research is advanced. Therefore, it can be said that LLMs are beginning to control physical systems and make decisions in the real world, a promising and rapidly coming into production direction in the future being that of building LLM agents, namely models that can autonomously carry out complex tasks. In the context of the research conducted to grant the LLMs direct control of virtual and physical systems, there have already been available some simple mechanisms among which the following can be mentioned: OpenAI's advanced data analyst tool that does not require the users to run the code themselves; LLMs equipped with web browsing capabilities that allow them to access websites and report back to the user; Google's PALM-E can take instructions as input and produce commands to control physical robots; ChatGPT, based on manually written code, can control drones with plain language descriptions; ChemCrow connects a LLM with an external scientific tool to issue control commands to robotic synthesis machines; LLMs agents have been successful in playing Minecraft; AI agents can schedule calendar, invites and browse the internet; AutoGPT is a scaffolding (related

to code built up around a LLM that allows it to use tools) software able to convert a LLM into an agent moving around the web (Woodside & Toner, 2024). Under these circumstances, new issues related to governance, accountability and ethics arise. They have been summarized as follows: LLM attack surface allow them to be exploited by malicious agents, liability in case of malfunctions that cause harm, disclosure of both agent and user, ability of AI agents/models to prevent risks generated by other AI systems, restrictions, regulations to mitigate the risks of out-of-control and power-seeking agents, which should be considered by developers, users and policymakers altogether (Woodside & Toner, 2024). In this context, mention should be made that now multi-agent systems (MAS) are considered, meaning teams of LLMs that can interact, without the need for a human to continually direct them, in order to solve complex tasks.

2. LLM-Related Challenges and Possible Solutions. Documented Cases

Having exposed the main general mechanisms through which LLMs can generate propaganda, either by being employed by malicious users or by their inherent functions and vulnerabilities, we consider appropriate to discuss some particular aspects related to the challenges they pose, especially to security in relation to propaganda, as they have been documented in the literature.

With regard to the LLM-generated propaganda persuasion capability, in the context of the growing concern, voiced by policymakers, technologists, and researchers, that the new AI tools could supercharge covert propaganda campaigns by allowing propagandists to mass produce text at low cost, in 2023, Stanford HAI conducted a survey experiment on 8,221 US respondents comparing the persuasiveness of English-language foreign covert propaganda articles sourced from real-world campaigns to text generated by a large language model. The topics were related to drones, Iran, the US–Mexico border wall, and the conflict in Syria. It was found that LLMs could create highly persuasive text, and that a person fluent in English could improve the persuasiveness of AI-generated propaganda with minimal effort. In figures, GPT-3-generated propaganda was highly persuasive, 43.5% of respondents agreeing with the thesis statement, compared to 24.4% in the control group, while the articles generated by GPT-3 with an edited prompt were as persuasive as the original propaganda (46.4% compared to 47.4%). Moreover, if a propagandist edited the input and selected the best of the three outputs on each topic, the GPT-3-generated propaganda would be even more persuasive than the original propaganda (52.7% compared to 47.4%) (Goldstein, 2024, p. 34). Considering the emerging GPT-4 as well as the visual and audio content (deepfake), it is obvious that propaganda can become increasingly effective.

In the context of the possibility to use LLMs to generate content that, as we have shown, proves to be credible, experts have debated the impact these tools can have on different fields of activity, especially following the launching of ChatGPT in November 2022. As far as journalism is concerned, three experts and two startup founders were thus invited by Reuters to express their ideas on the topic. The impact of AI becomes obvious even from the beginning, with reference to the interviewees' jobs, namely computational journalist, AI editor, Head of JournalismAI, a project by the London School of Economics journalism think tank Polis, as well as to the fact that the mentioned startup is an app that offers readers daily AI-generated brief summaries. Moreover, we find that there are many news publishers automating some content, including global agencies like Reuters, AFP and AP. In addition to automation, the development of LLMs offers new applications to journalism, journalists themselves testing the capabilities of chatbots to write and edit, feeling that there is an intelligence involved, even if it is still just a type of predictive technology, it is not original, and it does not have the required analytic capability, according to declarations (Adami, 2023). In this way and beyond the scope of the present paper, the LLM-generated content impact on other fields of activity, such as education and research can be analysed, again raising questions related to authorship and intellectual property.

To continue with the challenges to security that may be posed by LLMs, used as tools to generate automatically convincing and misleading text, besides the fact that AI tools can help malicious actors to spread disinformation and scale their operations, the phenomenon leads to legitimately questioning the percentage of truth in the content that is consumed online, as well the ways in which authenticity can be determined. Thus, in a report that was released to the public in early 2024, NewsGuard internet trust organization has identified as much as 725 unreliable websites publishing AI-generated news and information, which may be overseen by human agents or not, meaning that, in the absence of editorial control, the news and information did not meet journalistic standards (News Guard, 2024). Also, in early 2024, Google has advanced an experimental AI tool to be used by a selected group of independent publishers in the United States of America. Within this experiment, the beneficiaries had to publish daily three articles generated by AI, which leads to the conclusion that "traditional lines that enable trust in online content can be easily blurred" (Karanasios & Risius, 2024). The study conducted by two Queensland Professors emphasizes the fact that digital products, as they become essential on a large scale both for businesses and also in everyday life, "serve as a tool for platforms, AI companies and big tech to anticipate and push back against government", a possible explanation for the fact that the World Economic Forum's 2024 Global Risk Report predicts misinformation and disinformation to be the greatest threats for 2025-2026 (Karanasios & Risius, 2024).

On the other hand, to be as compliant as possible with the objectivity standard, mention should be made that LLM-generated propaganda is employed not only by malicious agents, which can be individuals or non-state actors, but also by state actors. In this context, a Freedom House report published in October 2023 in MIT Review shows that AI-generated texts are also used by governments and political actors around the world, in both democracies and autocracies, to manipulate public opinion in their favour and to automatically censor critical online content. Thus, researchers documented the use of generative AI in 16 countries to influence public debate. Moreover, it was found that, in 2023, global internet freedom declined for the 13th consecutive year, driven in part by the proliferation of artificial intelligence. In this context, some examples are provided, such as those of Venezuela, where pro-government messages were spread through AI-generated content. In addition, the report shows that a combination of human and bot campaigns was used to manipulate online discussions, at least 47 governments having deployed commentators to spread propaganda in 2023, which was double the number a decade ago. The cited report also discusses the normalization of the AI-generated content, which can make people more sceptical to true information, especially in times of crises and political conflict. Freedom House researchers documented 22 countries that passed laws requiring or incentivizing internet platforms to use machine learning to remove unfavourable online speech, China and India being examples in this case. In all, a record high of 41 governments blocked websites for political, social, and religious speech during the year covered by the study (Ryan-Mosley, 2023).

In the same vein, a research work conducted in 2023 by the University of Washington, Carnegie Mellon University and Xi'an Jiaotong University, presented at the Association for Computational Linguistics conference, consisting of testing 14 large language models on political biases, found that OpenAI's ChatGPT and GPT-4 were the most left-wing libertarian, while Meta's Llama was the most right-wing authoritarian. Feminism and democracy were among the addressed topics, models being retrained to detect hate speech and misinformation, potentially causing real harm. Some of the provided examples refer to some models refusing to offer information about abortion and contraception, expressing support for taxing certain social categories, such as the rich, overemphasizing social conservatism and hate speech targeting ethnic, religious, and sexual minorities. It is thus obvious that such messages can result in more, even extreme, social and political polarization, with harmful or uncontrollable consequences. Another interesting aspect emphasized by the mentioned research is that the process of training data helped to reinforce models biases even further. Considering to remove biased content from data sets in order to mitigate biases in language models, researchers have come to the conclusion that it is not enough, on the one hand, and that it is very hard to clean a vast database of biases, on the other hand. In other words, no language model can be entirely free from political biases (Heikkila, 2023).

Having presented some of the mechanisms of generating propaganda through the use of LLMs as well as some documented proofs of its existence, we will further review certain proposed solutions to counter the phenomenon that may have consequences for security, be they AI-aided or regulatory ones, besides education and critical thinking abilities enhancement (Lupu, 2023).

In relation to the models inherent vulnerabilities, some of the proposed solutions include the strict input validation and sanitization, meaning filtering out malicious or unexpected prompts that could initiate unauthorized request; security audits and configuration reviews to confirm that internal resources remain shielded from the LLM; network segmentation to isolate the LLMs from sensitive internal resources; monitoring and alerting in the event of unusual or unauthorized activities; least privilege access, both in terms of data and access; ethical guidelines and usage policies for the model, the main ethical concerns in the context of the present paper being those related to bias propagation and misinformation spread (Bright; Protect AI, 2023). All these aspects become much more important considering the results of a survey conducted among developers in May 2023. 70% of all respondents in the USA, Germany, India, UK and Northern Ireland were using or planning to use Al tools in their development process that year and those learning to code from online resources increased from 70% to 80% since the 2022 survey (Stack Overflow, 2023).

In the same vein, another proposed solution is that of providing LLMs with the ability to detect propagandistic textual spans, which is related to the above-presented annotation and reinforcement tools, be they RLHF or RLAIF. As RLHF is found costly and even faulty, RLAIF is taken into consideration in relation to its ability to optimize the content and mitigate propaganda-associated risks. We illustrate this proposed solution with a study that has explored the ability of ChatGPT-3 and GPT-4, compared to human annotators, to detect and label spans with propagandistic techniques, such as loaded language, revealing the GPT-4 great performance potential in all the three assigned roles, namely annotator, selector and consolidator, the highest score being in the case of consolidator, meaning that GPT-4 is learning from the initial annotations to perform better, closer to expert consolidators' level (Hasanain & Ahmad, 2024).

As for the rules and regulations in the field, we will discuss those adopted at the European Union level.

3. EU Regulatory Framework in the Field

Acknowledging that the Artificial Intelligence (AI) will have an enormous impact on the way people live and work, the European Union (EU) has developed some regulatory documents meant not only to achieve the goal of becoming a global hub in AI, but also to set out clear transparency and reporting obligations for any

company placing an AI system on the EU market as well as for companies whose system outputs are used within the EU.

Thus, in 2018, the Commission and Member States established a Coordinated Plan, namely a strategy in the field, which was reviewed in 2021. The key actions mentioned in the new strategy focus on setting enabling circumstances for development of AI and uptake within the EU, making it an environment in which excellent quality is found from the laboratory up to the market, and guaranteeing that AI "works for people and is a force for good in society" (European Commission, 2021). Moreover, the Digital Europe and Horizon Europe programs were planned.

The mentioned Coordinated Plan goes hand in hand with the Proposal for a Regulation on Artificial Intelligence. Referring to this important act, as it is the first worldwide legislation of such complexity regarding AI, the Commission released its proposal in April 2021, setting out a risk-based approach to regulation designed to grow trust in technology and ensure the safety and fundamental rights of both people and businesses. The proposal establishes a differentiated regulatory structure, prohibiting some uses of AI while setting severe norms for high-risk usage and lighter regulations for AI systems that do not pose such great risks.

According to the EU internal decision-making process, in June 2023, the Artificial Intelligence Act was amended (European Parliament, 2023), and was adopted in March 2024 by the Parliament, and on 21 May by the Council, being fully implementable two years after entry into force. Some parts of the Act will be applied sooner, like the interdiction of AI systems that pose unacceptable risks, codes of practice, and rules on general-purpose AI systems. Obligations for high-risk systems are applicable 36 months after its entry into force. Generative AI, such as ChatGPT, is not considered as having high-risk. However, it will have to obey the transparency conditions and EU copyright law by revealing that the content is AI-aided, preventing the model from producing illegal content, and publishing information relating to copyrighted data used for training. Moreover, high-impact general-purpose AI models that would be able to produce systemic risk, such as the more innovative AI model GPT-4, would have to undertake systematic assessments and any severe incidents will need to be conveyed to the European Commission. In addition, any content either created or altered with the help of AI – images, audio or video files (as, for instance, deepfakes) – will have to be clearly identified as AI generated, so that users can be conscious that they are exposed to such material.

In the same vein, the Artificial Intelligence Act addresses not only the issue of AI systems classification, but also the use of general-purpose AI (GPAI) models. As for the new governance architecture, several bodies are set up (AI Office, AI Board, advisory forum for stakeholders, etc.) and penalties are established. Transparency, safeguarding fundamental rights and measures in favour of innovation are also considered. The main takeaways can be briefly summarized as follows: the Act bans

AI systems involved in cognitive behavioural manipulation and social scoring within the EU; it prohibits AI usage for predictive policing based on profiling and systems that make use of biometric data to categorize people according to race, religion, or sexual orientation; for GPAI, the Act requires compliance with transparency requirements, exempting from these regulations the systems used exclusively for military, defence, and research purposes (Council of the EU, 2024).

Another document to be mentioned is the one developed by the European Parliamentary Research Service, entitled Generative AI and watermarking, to explain main generative AI functioning mechanisms and related concerns. Among them, the following are mentioned: unauthorized exploitation of datasets, openness to misuse, potentially leading to plagiarism, privacy issues, and AI hallucination phenomena. The document emphasizes that the need to differentiate AI-generated synthetic content from human content has become a key policy issue, showing that people are increasingly unable to detect AI-generated content. In this context, the document shows that a range of approaches are being tested to trace how AI content is generated and to document its provenance, which include content labelling, the use of automated fact-checking tools, forensic analysis, meaning the content examination for inconsistencies or anomalies that indicate manipulation, and watermarking techniques, which create a unique identifiable signature that is invisible to humans but algorithmically detectable and that can be traced back to the AI model, thus enabling the detection of AI-generated content and the identification of its provenance. Although watermarking still have its limitations and drawbacks, it would be an important step forward in coexisting with AI in a decent and ethical manner (European Parliamentary Research Service/EPRS, 2023).

Conclusions

Artificial Intelligence has become increasingly important for society, economy, as well as policy-making. The transformer architecture, developed in 2017, enables AI to be brought into the mainstream, thus having the potential of a general-purpose technology, namely one characterised by pervasiveness, improvement and innovation. Moreover, it suggests helping to lead to possible new discoveries in different fields, to reduce the cost of goods and services, to make work more efficient and effective, which can have huge societal implications. However, at least for the moment, it has demonstrated some limitations, especially with regard to generative AI, with emphasis on LLMs, which allow users to input a variety of prompts to generate new content, such as text, images, videos, sounds, codes, 3D designs, and other media, based on their training to predict outcomes.

Having explained the LLMs functioning principle and their potential to generate biased content or propaganda, thus posing challenges to security, as well and having

identified the main intrinsic and extrinsic vulnerabilities and the proposed solutions to mitigate risks, we would like to add some closing remarks, mainly focused on the possible best ways for the humans and such tools coexistence, as, once they emerge, they should be integrated in our daily lives. Firstly, we consider that, in the context in which they represent a social technological challenge, they should remain only tools that can augment human intelligence and not replace it. Secondly, it is the issue of trust in these tools that should be thoroughly addressed, in relation to their fairness, explicability, transparency, robustness and privacy. Thirdly, as the results they provide can include a false narrative, either directly, by the technologically inherent vulnerabilities, or indirectly, as a result of their poisoning or hijacking by malicious actors, the cost of indiscriminately employing such tools in the decision-making process should be carefully considered. That is why we strongly support the idea of education, audit and accountability in working with such AI tools as LLMs, hoping that more people will be sincerely involved in sharing their experience in the field, taking into account that generative AI has become an extremely important societal challenge.

As for the way in which the EU AI Act addresses some of the concerns presented in the paper, it is intended to cover issues related not only to technology but especially to the potential risks that may have serious consequences for the citizens, including propaganda-related ones. Thus, the regulation bans the use of AI-aided techniques to influence behaviour, exploit individual's and group's vulnerabilities or socially score people, stipulating, in Annex III, the high-risk AI systems in areas such as biometrics, critical infrastructure, education and vocational training, employment, access to essential public services and benefits, law enforcement, migration, asylum and border control management, administration of justice and democratic processes. Moreover, in relation to both AI systems and GPAI models, the Act requires compliance with transparency standards. However, some areas are mentioned as exceptions. Among them, we consider relevant to the topic of the paper the AI systems and models developed and used exclusively for military, defence and national security purposes, as well as those that are intended for scientific research and development.

BIBLIOGRAPHY:

Adami, M., 2023. *Is ChatGPT a threat or an opportunity for journalism? Five AI experts weigh in.* Available at: https://reutersinstitute.politics.ox.ac.uk/news/chatgpt-threat-or-opportunity-journalism-five-ai-experts-weigh

Bommasani, R. & Liang, P., 2021. *Reflections on Foundation Models,* s.l.: Stanford University Human-Centered Artificial Intelligence.

Bright; Protect AI, 2023. *Practical Guide. Exploring the Risks of Using Large Language Models,* s.l.: s.n.

Burtell, M. & Tonner, H., 2024. *The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1.* Available at: https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/

Council of the EU, 2024. *Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI,* s.l.: s.n.

Dahl, M. e. a., 2024. *Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive.* Available at: https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

European Commission, 2021. *Coordinated Plan on Artificial Intelligence 2021 Review*, s.l.: s.n.

European Parliament, 2023. *Artificial Intelligence Act,* s.l.: s.n.

European Parliamentary Research Service (EPRS), 2023. *Generative AI and watermarking.* Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf

Goldstein, J. e. a., 2024. How persuasive is AI-generated propaganda?. *PNAS Nexus,* 20 February.3(2).

Hasanain, M. & Fatema Ahmad, F. A., 2024. *Large Language Models for Propaganda Span Annotation.* Available at: https://arxiv.org/html/2311.09812v2

Heikkila, M., 2023. AI language models are rife with different political biases. *MIT Technology Review,* August.

IBM, 2024. *What are AI hallucinations?* Available at: https://www.ibm.com/topics/ai-hallucinations

Karanasios, S. & Risius, M., 2024. *Algorithms are pushing AI-generated falsehoods at an alarming rate. How do we stop this,* s.l.: s.n.

Lică, D., 2024. *The Terminological Conundrum Regarding Information Weaponisation.* Bucharest, CDSSS, Carol I National Defence University.

Lupu, D.-C., 2023. Critical Discourse Analysis in the Age of Multimodal Communication. *Romanian Military Thinking Journal,* Issue 3, pp. 184-199.

NATO, 2020. *Public Affairs Handbook.* Available at: https://www.act.nato.int/wp-content/uploads/2023/06/nato-pao-handbook-2020.pdf

News Guard, 2024. *Reports about online misinformation and disinformation from NewsGuard's analysts*. Available at: https://www.newsguardtech.com/reports/

Nvidia, 2024. *Large Language Models Explained.* Available at: https://www.nvidia.com/en-us/glossary/large-language-models/

Ryan-Mosley, T., 2023. How generative AI is boosting the spread of disinformation and propaganda. In a new report, Freedom House documents the ways governments are now using the tech to amplify censorship. *MIT Review,* October.

Stack Overflow, 2023. *2023 Developer Survey,* s.l.: s.n.
Vaswani, A. & al, e., 2017. *Attention is All You Need.* Long Beach, CA, USA, s.n.
Woodside, T. & Toner, H., 2024. *How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2.* Available at: https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/