Stance/Opinion/Consideration

Permanent Committee’s on Ethics and Regulation of Artificial Intelligence Stance on the Ethical Questions of Generative Artificial Intelligence and Large Language Models

Bratislava, 20th June 2023

The Permanent Committee on Ethics and Regulation of Artificial Intelligence (CERAI)

In the midst of a widespread societal debate surrounding the regulation of artificial intelligence (AI) systems which specifically use Large Language Models (LLMs), i.e. are built on Generative Artificial Intelligence technology, it is incumbent upon us, the Permanent Committee on Ethics and Regulation of Artificial Intelligence (CERAI), to present our joint stance containing a set of fundamental recommendations for the operators, regulators and users of these systems. A number of global initiatives, including open letters from AI scientists, and other high profile news in the mainstream media, gave an impetus for a discussion on the topic during our regular monthly meetings. Our stance was formed by the conclusions of this discussion. CERAI’s stance focuses on Generative AI. However, in the context of the debate on the proposed AI Act, some of the recommendations may be applicable to the development, deployment and use of AI in general. The AI Act was introduced by the European Commission in April 2021. In December 2022, the European Council published their general approach and after an adoption of the European Parliaments negotiating position in June 2023, the AI Act negotiations entered their trilogue stage. Based on CERAI’s per rollam vote, our committee recommends the following list of:

1. In the context of AI regulation, we need clear limits of the deployment and use of Generative AI systems. Furthermore, we need clarity of definition and scope of persons who are eligible to interact with these systems.
2. The operators of Generative AI systems should ensure the protection of persons, for whom their use of Generative AI represents disproportionate risk of negative impact on their life, health and safety. A priority focus is on children and other vulnerable groups. The identification of these vulnerable groups and the methods of their protection should be legislated.
3. The users of Generative AI systems must be informed that they are about to embark on an interaction with a non-human agent, i.e. they are about to be exposed to, process and consume the outputs generated by a machine. The operators must display such an explanatory information in a visible place, where the interaction with such a system is taking place throughout the duration of the interaction, together with a warning, that such a machine generated output may contain non-true and non-verified information.
4. The operators of Generative AI systems must make the information about the source and the nature of their training and validation data sets available to the regulators and notified bodies in accordance with the proposed AI Act, with trade secrets exemption provisions not applicable.
5. The operators of the Generative AI systems are obliged to communicate in an understandable manner how they process the user’s data and the interaction data of their users. In particular with regards the processing of the user’s private and sensitive data, so that the operators can verifiably demonstrate that they are in compliance with the existing legal requirements set out in the data protection regulatory regime already in force.

6. The operators of Generative AI systems must protect the intellectual property rights with regards the data used by these systems and data that was used to train these systems.
7. The operators of the Generative AI systems must act bona fide/fairly in their contractual dealings with the persons who participate in the development and deployment of these systems. The operators must provide them with fair and adequate conditions and appropriate remuneration for the delivered work.
8. The operators of Generative AI systems must ensure security of these systems according to the current highest achievable standards, including physical and cyber security. The operators are obliged to monitor and evaluate the attempts of misuse of these systems and implement targeted measures aimed at their elimination.
9. The operators of Generative AI systems, as well as public institutions, adopt a responsible and methodical approach to the education and continuous updated information sharing with their users, as well as the general public in the context of the use of Generative AI systems, so that the users and the general public achieve and maintain accurate knowledge of the advantages, but also the limits and risks associated with the use of these systems, as they keep evolving in real time.
10. The operators of Generative AI systems must demonstrate an adequate effort in ensuring the diversity, objectivity and credibility of the generated output for the users.
11. The regulation of Generative AI systems should fall within the high risk category, taking into consideration the application area, the purpose of such systems, and also the degree of the risks of their overall societal impact.
12. The operators of Generative AI systems must undergo a conformity assessment when entering the market or deploying such a system into production under the supervision of the notified bodies in accordance with the proposed AI Act.
13. The operators of Generative AI systems facilitate cooperation with the regulatory bodies and notified bodies as per AI Act during the independent audits of their systems and regularly monitor their systems with the aim of minimizing/mitigating the potential risks even after their market entry or deployment into production.

Explanatory report

In the past few years, we have witnessed an ascend/proliferation of AI systems, which are built on the technology of Large Language Models (LLMs), i.e. using the principles of generative AI for various modalities of the output, e.g. creation of texts, images, human voice, or other synthetically generated outputs. AI systems became ubiquitous. The LLM technology dates back decades, however, the applied innovations specifically in the past decade, i.e. since 2012, such as the publication of a paper on „dropout“, which limits a bugbear of neural network training known as „overfitting“, as well, as the 2017 paper on „Attention mechanism“[1], together with the massive volumes of centralized data (big data) and exponential increase in the computing capacity (resulting in huge quantities of energy-sucking GPUs for training and use) enabled the operators of Generative AI systems to create tools, which are capable of mass-generation of outputs comparable with the creative capabilities of humans, i.e. the outputs are of such high quality that for a layman they are indistinguishable from the outputs generated by humans. By deploying individual instances of Generative AI systems into production, such as ChatGPT, DALL-E, Midjourney, the global general population was exposed to the mass use of these tools. The general population had no prior experience with such tools. No prior

[1] Back to the Fifties: Reassessing Technological and Political Progress - American Affairs Journal

knowledge about how they work. No prior knowledge about their capabilities, nor their limitations and potentially harmful risks.

CERAI is cognitive of the fact, that the use of the tools of Generative AI or tools based on LLMs, brings a lot of advantages for their users. We see and acknowledge the attempts to improve the educational methods, health care or their use in the design, business workflows and marketing.[2] On the other hand, we are aware of the risks that the mass usage of these tools brings about. We are receptive to various initiatives, such as the open letter of the Future of Life Institute (FLI) [3], and appreciate the efforts of well-known signatories to stoke a cross-societal debate not only about the advantages and the benefits, but also about the potential risks associated with the use of the Generative AI systems and LLMs. We also monitor the reported instances of materialized impact of the aforementioned risks, such as life threatening situations, risks to human health and human rights.

CERAI holds an opinion, that at this point in time, an urgent re-focus of the attention of the general public on the current societal issues of the use of Generative AI systems of short to medium term horizon is required. We are concerned with the mainstream media sensational presentation of the potential threats and risks, e.g. as mentioned in the open letter of the FLI, such as the creation of artificial minds, the loss of the meaning of life, the eradication of jobs, or the extinction of human race, as distracting from the „matter of fact“ threats and risks of our presence and immediate future. This hyped up presentation of sensational threats and risks has not only captivated the imagination of general public, but has a potential to command the attention of our political elites, at the expense of substantially more pressing „matter of fact“ threats and risks we are facing now and in our near future.

In this context, CERAI is of a conviction, that the operators of the tools of Generative AI should, as all the other operators of AI systems, follow the principles of the development and the deployment of trustworthy and ethical AI as presented by the European Commission's[4] Expert group or other international organizations[5], and be compliant with the existing legal frameworks with a consideration of the specific requirements, which are characteristic for these technologies. In this aspect, we rate the requirements of transparency, privacy, fairness and human oversight, as of the utmost importance and urgency.

One of the most serious issues challenging our societies is the conscious and/or unconscious anthropomorphism of the tools of Generative AI not only by their users, but also by the deliberate marketing strategies of these tools by their operators. We are aware that this issue is not a brand new issue. The attempts to anthropomorphize artificial systems have been present historically whenever deploying systems which imitate some of our human features and behaviors. However, the machine outputs which started to imitate convincingly human expressions or to generate unverified and untrue static or dynamic likeness of real events or humans in certain situations which are indistinguishable by humans from the actual rendition of the real events and humans[6] in certain situations, cause a grave concern to us. These

[2] https://research.aimultiple.com/generative-ai-applications/

[3] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[4] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 alebo https://unesdoc.unesco.org/ark:/48223/pf0000381137

[6] https://www.theverge.com/2023/3/27/23657927/ai-pope-image-fake-midjourney-computer-generated-aesthetic

outputs substantially undermine our human autonomy and our ability to form our opinions freely, even if only temporarily*.* *(Winston Churchill quote: „A lie gets halfway around the world before the truth has a chance to get its pants on".)* The potential risks and harms to human life of these Generative AI systems are materializing in front of our eyes, e.g. reported instances of people from vulnerable groups or people in difficult life situations, resorting to desperate actions, in some documented cases even suicides[7] after a communication with human like chatbots. Some users are willing to change major life decisions and even their moral attitudes based on the outputs generated by AI systems[8], and furthermore, even denying that such an influence or an intervention into their decision making autonomy took place. It is becoming clear that Generative AI systems have the ability to influence major life event decision making of humans that interact with these systems, without humans consciously realizing it. This ability of generative AI systems creates an opportunity for mass generation and spreading of disinformation online, manipulation of political views and public opinion and undermining public trust in our democratic institutions.

The above materialized events of the risks and harms of Generative AI systems led CERAI to a conclusion that the operators of the Generative AI systems should not create a false/fake impression that the users are communicating with a human being, i.e. an entity with a conscience, or an entity capable of compassion. From the standpoint of AI regulation, **we need clear limits of the deployment and use of Generative AI systems. Furthermore, we need clarity of definition and scope of persons who are eligible to interact with these systems.** Therefore it follows, that **the operators of Generative AI systems should ensure the protection of persons, for whom their use of Generative AI represents disproportionate risk of negative impact on their life, health and safety. A priority focus is on children and other vulnerable groups. The identification of these vulnerable groups and the methods of their protection should be legislated.** CERAI is of the opinion, that **the users of Generative AI systems should be informed that they are about to embark on an interaction with a non-human agent, i.e. they are about to start processing and consuming machine generated output. The operators are obliged to provide this explanatory information in a place visible throughout the duration of the interaction, together with a warning, that machine generated output might contain non-true and non-verified information.** This is recommendation is in line with the requirement of transparency of AI systems as defined in the proposed AI Act. The first version of the AI Act proposal was published in April 2021[9]. On the other hand, CERA is of a view, that there is also a requirement on a responsible use of Generative AI systems by the users. In case, that users are ignoring the guidelines, not exercising due care either knowingly or due to negligence when interacting with Generative AI systems, the users should assume their share of responsibility for the negative impact of such actions from their side.

Another big problem of generative AI is the issue of privacy and data governance. It concerns not only a possible leak of sensitive data[10], but also a transparent publication of the origin of the training and validation data sets of Generative AI systems. We are aware of publicly documented instances, e.g. OpenAI chose not to publish the information about the origins of

[7] https://cybernews.com/news/man-takes-own-life-chatbot/

[8] https://www.nature.com/articles/s41598-023-31341-0

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[10] https://www.engadget.com/openai-says-a-bug-leaked-sensitive-chatgpt-user-data-165439848.html

the training data sets[11] for their LLM GPT-4. By choosing not to be transparent about the origins of the training data set, they management of OpenAI effectively disabled audit and control of whether they are using data without prior consent of concerned persons. Journalists, AI experts but also some regulatory bodies[12] are alerting to this fact. CERAI is of the opinion, that **the operators of Generative AI systems should make the information about the origin and the nature of their training and validation data sets public to regulators and the notified bodies as per AI Act, with trade secrets exemption provisions not applicable.** At the same time, **the operators of the Generative AI systems are obliged to communicate in an understandable manner how they process the user's data and the interaction data of their users. In particular with regards the processing of the user's private and sensitive data, so that the operators can verifiably demonstrate that they are in compliance with the existing legal requirements set out in the data protection regulatory regime already in force.** On the other hand, the users should behave responsibly, and be aware that the communication/interaction with the chatbot/machine is not private. If the users are willing to share their private and sensitive information[13], despite this warning, they should assume a proportionate share of their responsibility for their knowing and willing actions of sharing their private and sensitive information.

The operators of Generative AI systems should ensure the protection of IP rights with regards the data that are being used by their systems, as well as the data used to train their systems. CERAI is of the conviction, that the protection of IP rights is key. We are aware that the implementation of this goal represents a substantial challenge, given that the training data sets come from various heterogeneous sources, and it can be difficult to follow the paths. Therefore, it is of utmost importance to implement procedures/processes which will enable the protection of IP rights but will not stifle innovation. These guidelines are embedded in the requirements of transparency and the notification obligation when using IP protected materials in the training data sets. It is important to invest and support the development and use of technologies which automate the detection and governance of IP rights in the outputs of Generative AI systems, and thereby streamlining the audit process and making the compliance with these requirements feasible. CERAI supports an open dialogue among the operators of Generative AI systems, the IP rights owners/beneficiaries and regulatory bodies, so that these stakeholders jointly determine the best procedures and norms, which respect the rights of all concerned stakeholders/parties and support innovation in the area of AI.

CERAI would like to highlight also the risk, that the distribution of the benefits of the use of tools such as Generative AI will not always be just or fair. Generative AI systems and LLMs might demonstrate and enhance, just like other AI systems, existing societal biases and discriminatory practices[14]. Generative AI systems may, just like other AI systems[15], present various forms of biases, they might under-represent or insufficiently present not only minority views, but also societally prevalent views[16]. In the past, we also witnessed how some AI

[11] https://www.siliconrepublic.com/machines/openai-gpt4-transparency-ai-concerns-stripe-chatgpt

[12] https://www.bbc.com/news/technology-65139406

[13] https://www.darkreading.com/vulnerabilities-threats/samsung-engineers-sensitive-data-chatgpt-warnings-ai-use-workplace

[14] https://www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/

[15] https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/

[16] https://arxiv.org/pdf/2303.17548.pdf

systems designed for communication expressed and enhanced racial and religious intolerance[17]. In our present, the operators are aware of this, and are trying to mitigate such biased, intolerant output, however, often unfortunately at the expense that some group of population of the global south are exposed to excessive unwanted/undesired content and used as a cheap labor to moderate and control and label such outputs[18] generated by such AI systems. On one hand, CERAI is of the opinion that the operators of Generative AI systems should try to mitigate the occurrence of societal biases and discriminatory expressions in the outputs of these systems. On the other hand, CERAI is of the opinion, that **the operators of the Generative AI systems must act bona fide/fairly in their contractual dealings with the persons who participate in the development and deployment of these systems. The operators must provide them with fair and adequate conditions and appropriate remuneration for the delivered work.**[19]

As per the aforementioned points, CERAI would like to express our view, that Generative AI systems or LLMs need to be regulated on the side of their development and deployment, but also on the side of a responsible user. We consider the proposed ban or temporary ban on the training of LLMs, like mentioned in the open letter of LFI, are not an effective solution. We identify ourselves with a view that our focus and resources should not be solely invested in the further development of ever larger and more powerful tools of Generative AI systems, but also the research of their security aspects and their impact on the various areas of society. **The operators of Generative AI systems must ensure security of these systems according to the highest current achievable standards, including physical and cyber security. The operators are obliged to monitor and evaluate the attempts of misuse of these systems and implement targeted measures aimed at their elimination.**

The operators of Generative AI systems, as well as public institutions, adopt a responsible and methodical approach to the education and continuous updated information sharing with their users, as well as the general public in the context of the use of Generative AI systems, so that the users and the general public achieve and maintain accurate knowledge of the advantages, but also the limits and risks associated with the use of these systems, as they keep evolving in real time. This fact is visible, for example when deploying these Generative AI systems in order to create and disseminate disinformation and harmful content, which impacts a large number of general population. **The operators of Generative AI systems must demonstrate adequate effort in ensuring the diversity, objectivity and credibility of the generated output for the users.**

CERAI asserts that the development, deployment and use of Generative AI systems should be the subject of legally binding regulation, such as the proposed AI Act. **The regulation of Generative AI systems should fall within the high risk category, taking into consideration the application area, the purpose of such systems, and also the degree of the risks of their overall societal impact.** In case, that the deployment of these systems falls into a prohibited category, such as the use of subliminal techniques, or the misuse of vulnerable areas with the intent of causing physical or psychological harm, then also such Generative AI systems, satisfying these conditions of prohibited category according to the AI Act, should also be prohibited. Therefore, in this context, CERAI is of the opinion, that **the operators of Generative AI systems must undergo a conformity assessment when**

[17] https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

[18] https://time.com/6247678/openai-chatgpt-kenya-workers/

[19] https://dl.acm.org/doi/10.1145/3461702.3462572

entering the market or deploying such a system into production under the supervision of the notified bodies in accordance with the proposed AI Act[20], just like the proposed requirement for the operators of the AI systems purposed for the use of remote biometric identification of persons. These notified bodies in accordance with the AI Act should be established by competent controlling bodies with the focus on their independence, expert capability and zero conflict of interests. **The operators of Generative AI systems facilitate cooperation with the regulatory bodies and notified bodies as per AI Act during the independent audits of their systems and regularly monitor their systems with the aim of minimizing the potential risks even after their market entry or deployment into production.**

[20] https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=CELEX:52021PC0206 (Kapitola 4)