Beyond the Hype: Toward a Concrete Adoption of the Fair and Responsible Use of AI

Lelio Campanile[1,*,†], Roberta De Fazio[1,†], Michele Di Giovanni[1,†] and Fiammetta Marulli[1,†]

[1]Department of Mathematics and Physics Università degli Studi della Campania "L. Vanvitelli", viale Lincoln 5, Caserta, 81100, Italy

Abstract

Artificial Intelligence (AI) is a fast-changing technology that is having a profound impact on our society, from education to industry. Its applications cover a wide range of areas, such as medicine, military, engineering and research. The emergence of AI and Generative AI have significant potential to transform society, but they also raise concerns about transparency, privacy, ownership, fair use, reliability, and ethical considerations. The Generative AI adds complexity to the existing problems of AI due to its ability to create machine-generated data that is barely distinguishable from human-generated data. Bringing to the forefront the issue of responsible and fair use of AI. The security, safety and privacy implications are enormous, and the risks associated with inappropriate use of these technologies are real. Although some governments, such as the European Union and the United States, have begun to address the problem with recommendations and proposed regulations, it is probably not enough. Regulatory compliance should be seen as a starting point in a continuous process of improving the ethical procedures and privacy risk assessment of AI systems. The need to have a baseline to manage the process of creating an AI system even from an ethics and privacy perspective becomes progressively more important In this study, we discuss the ethical implications of these advances and propose a conceptual framework for the responsible, fair, and safe use of AI.

Keywords

Artificial Intelligence, Generative AI, Ethical AI, Large Language Models

1. Introduction

Artificial Intelligence (AI) is a rapidly advancing field of science and technology that has the potential to revolutionize various sectors of industry and society. With its ability to process vast amounts of data, generate insights, and support decision-making, AI has emerged as an important part of many organizations' processes. However, concerns about the impact of AI on society, particularly from an ethical perspective, have increased as its use has grown. From self-driving cars to virtual assistants, the applications of AI are endless as the quality and performance of AI techniques and methods continue to improve.

The advent of generative AI expands the potential applications of AI and increases the dangers it poses. Generative AI is a subset of AI that uses Machine Learning (ML) algorithms to generate new content based on existing data. It makes it possible to create content that appears as new and original, but is the result of generating statistics based on training data sets.

Generative AI raises new ethical challenges and a whole new set of emerging issues because of the difficulty in separating human-generated content from machine-generated content.

It becomes crucial a fair use of AI in any field of application, first and foremost in sensitive fields such as medical, military, and engineering, where the human decision-making component is of primary importance, but also in research and education where fair use of AI becomes critical to the informed growth of students with critical thinking and quality research. With the rapid developments in machine learning and generative AI models, the newborn of more powerful Large Language Model (LLM) models such as ChatGPT, Claude, Mistral and others continue to receive attention focusing on the associated risks, particularly from legal and ethical points of view.

There are both exciting opportunities and significant ethical challenges associated with the use of generative AI. The technology has the potential to revolutionize various sectors of society. However, it also raises concerns about job displacement, transparency, privacy, ownership, inequality, and reliability. To ensure that, the benefits of generative AI are maximized while its risks are minimized, the development of responsible and ethical frameworks for its use will be critical.

In this paper, we explore the key ethical issues, promises, and perils of AI use, and propose a conceptual framework that could contribute to the responsible, reliable, fair, and safe use of AI.

The rest of this paper is structured as follows: Section 2 gives a brief overview of AI and generative AI, Section

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

†These authors contributed equally.

lelio.campanile@unicampania.it (L. Campanile);
roberta.defazio@unicampania.it (R. D. Fazio);
michele.digiovanni@unicampania.it (M. D. Giovanni);
fiammetta.marulli@unicampania.it (F. Marulli)

0000-0003-4021-4137 (L. Campanile); 0000-0002-0271-132X (R. D. Fazio)

3 focuses on the ethical implications and issues of AI. Section 4 presents the conceptual framework. Finally, Section 5 presents the conclusion and future research directions.

2. AI and Generative AI background

In the last few years, Artificial Intelligence Generated Content (AIGC)[1] has gained outstanding popularity, not only in the computer science research community but, mainly in terms of interest in the various content generation products built by large tech companies. AIGC refers typically to contents that can be automatically generated by adopting advanced Generative AI (GAI) techniques, as opposed to being created by human authors.

GAI-based systems can automate the creation of large quantity of content in a very short time. The most representative exemplars are provided by the OpenAI tools, namely ChatGPT[2] and DALL-E[3]: these tools can generate, respectively, but not limited to, textual documents and pictures, exploiting large knowledge bases laying under the interaction systems, typically provided as conversational agents. The extraordinary popularity of these tools can be reasonably found exactly in the key aspects to being friendly and ready-to-use tools for not expert people: by adopting a very familiar interface, provided in the shape of an instant messaging system, properly called conversational agents or shortly as chatbots, common users are enabled to test and exploit effectively the potential of the generative technologies. ChatGPT is a Large Language Model (LLM)[4] — based tool, developed by OpenAI for building conversational AI systems, which can efficiently understand and respond to human language inputs in a meaningful way[5]. In addition, DALL-E is another state-of-the-art GAI model also developed by OpenAI, which is capable of creating unique and high-quality images from textual descriptions in a few minutes, such as "a pink rabbit going to Mars boarding its flying basket", in a photorealistic style. Anyway, GAI is not free from research challenges, concerning, for example, the appropriate set of commonly used evaluation metrics for assessing fidelity, faithfulness and quality of artificially generated data, as discussed in [6]. A further analysis concerning GAI methodologies and research aspects, along with a comprehensive classification of input and output formats used in GAI systems, is provided in [7].

Whether GAI represents a significantly challenging issue for researchers, involved in understanding and improving the representation of the knowledge that is behind, GAI-based systems are also carriers of not trivial implications, as the ones represented by social impact and related to ethical and legal aspects. Observing the phenomenon of the outstanding popularity of these kinds of systems and tools among common users, it brings back to mind the effects of Web 2.0, with the introduction of User Generated Contents (UGC)[8], where people were enabled to write everything almost everywhere. A deleterious phenomenon deriving from the exceeding democracy of the web still remains represented by the fake news unconditioned spreading[9], as discussed in the studies proposed in[10], [11], [12]. Fake news could be automatically generated by GAI systems, with features that make them challenging to be distinguished from real news when automatic classification systems are employed. With the very recent advances of GAI, generating fake content is within everyone's reach. Finally, also novel cyber-security issues are introduced by the malicious exploitation of generative AI[13]. Foremost among them there are the adversarial attacks, performed mostly by re-shaping and re-arranging well-known malicious behaviours and activities under a novel unknown guise, to cheat defence and intrusion detection systems[14]. Zero-days attacks, along with data and model poisoning attacks, are very frequently supported by GAI-based systems[15]. In [10] and [16] poisoning attacks targeting machine learning models, performed by the exploitation of adversarial and GAI are discussed. In the work of [17], a case study for energy distribution and dispatching systems frauds is discussed, highlighting the potential drawbacks and threats deriving from a malicious-aim driven exploitation of Generative Adversarial Networks (GANs)[18], several years before the current explosion of popularity of current GAI systems.

3. Ethics aspects: promises and perils

The emerging new possibilities associated with AI and GAI raise various ethical challenges that should be addressed in a comprehensive manner. Researchers, physicists, and engineers should not remain at the legal minimum compliance in terms of facing ethical issues in the field of AI, but they should study, understand, and act in a better possible way to mitigate or eliminate those issues.

A significant concern in AI is handled by bias. Starting from data collection to model training, bias is a potential risk in different stages of the AI process. The potential risk is to perpetuate the existing bias in training data to AI model. This risk becomes very high in GAI model training. [19]

In GAI, the amount of data used to train the model is enormous. Often this data has been collected from the Internet and using different and heterogeneous data sources. Unlike a traditional AI model, which is used

Figure 1: OpenAI Dall-E-2 photorealistic image

for prediction or classification purpose, the generative model is used to create new content. Bias in training data increases, in this case a specific ethical issue, perpetuating a potential social bias in the newly generated content.

In addition, researchers working in this area face another ethical dilemma: if the data contain biases that reflect society, is it correct to work to mitigate these biases? If so, how?

Certainly, it must be done with the utmost care, because biased AI systems could potentially exacerbate existing societal inequalities. They could perpetuate prejudice or reinforce stereotypes. They could also produce disparate outcomes for groups based on factors like race, gender, or socioeconomic status, leading to further inequality and social unrest. There is a real risk of perpetuating harmful stereotypes and possibly even distorting beliefs [20].

Strictly related to the bias issue, in the special mode with generative content creation, there is the problem of misleading information and fake news generation.

The ability of LLMs to create information that is not present in their original data, known as hallucination of LLMs or more technically called emerging feature [21], introduces the problem of creating misleading text, which could easily become fake news.

Moreover, recent developments in GAI allow not only text, but also images (figure 1), video, and audio to be created, enabling non-technical people to effectively use these techniques through simple applications.

It is clear that illegal use of these technologies has led to attempts at fraud and extortion, and can also lead to major legal and social problems. The illegal use of these techniques to create images and videos that substitute the face or other physical characteristics of one person for those of another. Because of their potential to create believable and deceptive content that can be used to spread misinformation or damage the reputation of individuals. The most relevant privacy issues include:

- Privacy Violation: fake content can be used to manipulate existing videos or images without the consent of the people involved, possibly violating their privacy.
- Identity Theft: by spreading false information, misleading content, or malicious messages, fake content can be used to impersonate individuals and cause significant damage to their reputations and privacy, as well as organize financial fraud.
- Revenge Porn: fake content can be used to create not real videos or images that show people in compromising situations, damaging their privacy and reputation. In the most serious cases, money is solicited for extensive purposes.
- Misinformation or Disinformation: fake content can have a significant impact on public opinion, trust, and decision-making by spreading false information or propaganda. This misinformation can also have a serious impact on society. It can lead to social unrest, political instability, and other negative consequences.

It is important to emphasize that privacy issues arise first in AI processes. In fact, there are significant privacy issues at the data collection stage, because in this stage is where sensitive information is collected and stored, making it vulnerable to potential security breaches and unauthorized access. [22], [23]

Finally, it is also interesting for this discussion to mention the problem of copyrighting the content on which AI systems, especially GAI systems, are trained. Often the source of this data is not really known.

GAI systems can use, process, and generate content without explicit consent, potentially violating the privacy of individuals and organizations.

The considerations presented here on the ethical risks associated with AI and the perils that arise from them, depend in great part on unaccountable or unfair use of AI, both by the creators of AI systems and by the end users of such technologies.

In the field of text generation, there are a lot of use cases where LLMs can help and improve the regular activities of students and researchers if it is used in a fair way.

GAI systems, such as ChatGPT, could be leveraged for students to get ideas or insights on specific topics. If you know well the idea that you want to write, then the GAI could help you to write it without grammar mistakes, especially if you are writing in non-native language.

This could greatly benefit non-native speakers, even in academia, in a sort of democratization of the dissemination of scientific thought, without having to resort to expensive language revision services.

On the other hand, an unfair and unethical use of this technology by students and researchers raises a very important ethical and legal problem related to authorship.

The need to know whether a piece of content is human-generated or machine-generated is becoming more relevant and critical.

4. A conceptual framework

Faced with these ethical and practical problems, the governments of various countries around the world have not stood still.

The United States has responded with the AI Bill of Right [24], which is not a regulation but only a white paper of recommendation from the White House Office of Science and Technology Policy. It outlines the main principles to be followed to pursue ethical issues in AI. It is a guideline for designing and deploying AI systems that respect human rights, enhance fairness, and protect personal privacy.

The European Commission has gone further with the EU AI Act, a "Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence and amending certain Union legislative acts" [25]. It is a fully-fledged legislative proposal that aims to address the risks associated with artificial intelligence systems.

The AI Act intends to ensure that AI systems are trustworthy, reliable, and beneficial to individuals and society.

Furthermore, once again the European Commission enacted the General Data Protection Regulation (GDPR) [26], which although not closely related to artificial intelligence, protects the privacy rights of European citizens, with particular emphasis on the automatic gathering and processing of personal data.

These documents provide a solid foundation for the development of a conceptual framework to assist researchers and companies in developing and deploying AI systems that are not only compliant but also address and attempt to solve AI ethical issues.

The four pillars on which the proposed framework is built are:

- Explainability Artificial Intelligence (XAI)
- Use of tools when possible
- Audit and organization for ethical compliance
- Continuous risk assessment

In figure 2 is depicted a possible workflow for the application of this framework.

The trustworthiness and transparency of an AI system are important characteristics for the responsible use of AI, because they increase the sense of security in using the system and the confidence in it. XAI is a cornerstone to achieve these aims. The user should be able to understand why the AI system arrives at the results it does and why certain actions are taken. XAI supports transparency, which not only increases user confidence, but also helps organizations analyse potential biases and errors in their models.

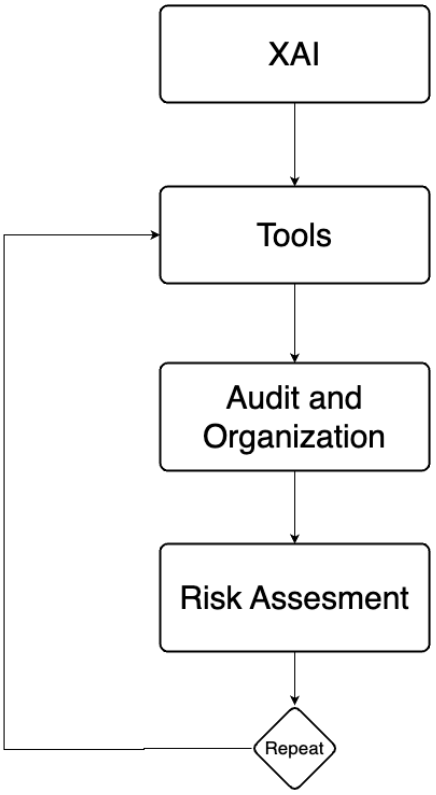

Figure 2: Workflow for the application of the conceptual framework for responsible and ethical use and development of AI Systems

To be effective, the application of the framework should be continuous in time and cyclical. The adoption, where possible, of various tools and techniques to review the fair use of AI systems will be essential. These should include automatic systems to check the origin of training data, or tools that can help assess whether a text is human-generated or machine-generated. Tools that protect against adversarial attacks and data poisoning attacks are also needed to keep the system fair, ethical, and secure.

In this field, the importance of research is paramount because even though some steps have been taken in the right direction, new developments are moving fast, and it is necessary to constantly improve tools and techniques.

The next phase involves regular audits and organizational practices to encourage ethical and responsible use and development of AI systems.

This could include internal reviews of development processes, ongoing training for operators, and conducting regular audits to assess the ethical implications of AI

systems. These practices should be organized with clear guidelines to avoid any misunderstanding or abuse of AI techniques.

Finally, a regular and cyclical risk assessment process specific to AI systems is required to promptly identify, evaluate, and prioritize potential risks associated with the development of AI systems.

5. Conclusion and Future Works

Generative AI is all about creating artificial data that looks like the real thing. This super-realistic data can be a game-changer in many fields, from video games to medicine and finance, until the arts. The resulting production of GAI is sometimes referred to as "fake data", to evidence that the contents were generated by an automatic process performed by a machine and not by a human being. GAI enables to generate fake but realistic images, to write new text, compose music, and even build chatbots that seem like chatting with real people. Besides the research efforts to improve the quality of the AI production, several ethical, legal and security issues need to be addressed.

It is apparent to need to address these issues systematically and beyond mere regulatory compliance. The development of a conceptual framework to address these issues should be a good starting point.

Future work will include improving the framework and exploring ways to make it more practical, including measures of the performance of ethical and responsible use of AI and GAI.

Moreover, we plan an in-depth look at the topic of detecting human user-generated texts from texts generated by a GAI, exploring existing techniques and towards new ones. Finally, we will continue research in the area of XAI (whose exploration began in[27] and [28]), which also extends to GAI, in order to improve the transparency of AI systems.

References

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, L. Sun, A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt, arXiv preprint arXiv:2303.04226 (2023).

[2] OpenAI, Conversation with chatgpt, 2023. URL: https://chat.openai.com.

[3] OpenAI, Dall-e 2, 2023. Generative AI model for image creation. https://openai.com/dall-e-2.

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology (2023).

[5] M. Abdullah, A. Madain, Y. Jararweh, Chatgpt: Fundamentals, applications and social impacts, in: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), Ieee, 2022, pp. 1–8.

[6] F. Marulli, P. Paganini, F. Lancellotti, Exploring the faithfulness of synthetic data by generative models, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 2214–2221.

[7] A. Bandi, P. V. S. R. Adapa, Y. E. V. P. K. Kuchi, The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges, Future Internet 15 (2023) 260.

[8] B. Omar, W. Dequan, Watch, share or create: The influence of personality traits and user motivation on tiktok mobile video usage (2020).

[9] F. Marulli, S. Marrone, L. Verde, Sensitivity of machine learning approaches to fake and untrusted data in healthcare domain, Journal of Sensor and Actuator Networks 11 (2022) 21.

[10] F. Marulli, L. Verde, L. Campanile, Exploring data and model poisoning attacks to deep learning-based nlp systems, Procedia Computer Science 192 (2021) 3570–3579.

[11] F. Marulli, L. Verde, S. Marrore, L. Campanile, A federated consensus-based model for enhancing fake news and misleading information debunking, in: Intelligent Decision Technologies: Proceedings of the 14th KES-IDT 2022 Conference, Springer, 2022, pp. 587–596.

[12] L. Campanile, P. Cantiello, M. Iacono, F. Marulli, M. Mastroianni, Vulnerabilities assessment of deep learning-based fake news checker under poisoning attacks, Computational Data and Social Networks (2021) 385.

[13] L. Campanile, M. Iacono, F. Martinelli, F. Marulli, M. Mastroianni, F. Mercaldo, A. Santone, Towards the use of generative adversarial neural networks to attack online resources, in: Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 34th International Conference on Advanced Information Networking and Applications (WAINA-2020), Springer, 2020, pp. 890–901.

[14] O. Eigner, S. Eresheim, P. Kieseberg, L. D. Klausner, M. Pirker, T. Priebe, S. Tjoa, F. Marulli, F. Mercaldo, Towards resilient artificial intelligence: Survey and research issues, in: 2021 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2021, pp. 536–542.

[15] C. A. Visaggio, F. Marulli, S. Laudanna, B. La Zazzera, A. Pirozzi, A comparative study of adversarial attacks to malware detectors based on deep learning, Malware Analysis Using Artificial Intelligence and Deep Learning (2021)

477–511.

[16] L. Verde, F. Marulli, S. Marrone, Exploring the impact of data poisoning attacks on machine learning model reliability, Procedia Computer Science 192 (2021) 2624–2632.

[17] F. Marulli, C. A. Visaggio, Adversarial deep learning for energy management in buildings, in: Proceedings of the 2019 Summer Simulation Conference, 2019, pp. 1–11.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.

[19] K. Wach, C. D. Duong, J. Ejdys, R. Kazlauskaitė, P. Korzynski, G. Mazurek, J. Paliszkiewicz, E. Ziemba, The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt (2023). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85183620669&doi=10.15678%2fEBER.2023.110201&partnerID=40&md5=deab98413c32b948ba57308e7e53fa6a. doi:10.15678/EBER.2023.110201.

[20] M. Zhou, V. Abhishek, T. Derdenger, J. Kim, K. Srinivasan, Bias in generative ai, arXiv preprint arXiv:2403.02726 (2024).

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions (2023). arXiv:2311.05232.

[22] Y. Zhang, M. Wu, G. Y. Tian, G. Zhang, J. Lu, Ethics and privacy of artificial intelligence: Understandings from bibliometrics, Knowledge-Based Systems 222 (2021) 106994.

[23] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, Z. Lin, When machine learning meets privacy: A survey and outlook, ACM Computing Surveys (CSUR) 54 (2021) 1–36.

[24] White House Office of Science and Technology Policy, Ai bill of right, 2022. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights, accessed on 01 April, 2024.

[25] Council of European Union, Eu ai act, 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, accessed on 01 April, 2024.

[26] Council of European Union, Regulation (eu) 2016/679 of the european parliament and of the council - general data protection regulation, 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504, accessed on 01 April, 2024.

[27] L. P. Di Bonito, L. Campanile, E. Napolitano, M. Iacono, A. Portolano, F. Di Natale, Analysis of a marine scrubber operation with a combined analytical/ai-based method, Chemical Engineering Research and Design (2023). doi:https://doi.org/10.1016/j.cherd.2023.06.006.

[28] L. Campanile, L. Di Bonito, M. Iacono, F. Di Natale, et al., Prediction of chemical plants operating performances: a machine learning approach, PROCEEDINGS EUROPEAN COUNCIL FOR MODELLING AND SIMULATION 2023 (2023) 575–581.