

From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI

Valentina Franzoni[1,2(✉)]

[1] Department of Mathematics and Computer Science, University of Perugia, PG 06123 Perugia, Italy
[2] Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong, China
valentina.franzoni@unipg.it

Abstract. The rapid development of Artificial Intelligence (AI) systems has raised significant ethical concerns, particularly with the problem of transparency in their decision-making processes. As AI systems become increasingly integrated into various aspects of society, there is an urgent need to transform these 'black box' models into more transparent and understandable 'glass-box' systems. This paper explores the methods, challenges, and implications associated with supplementing transparency in AI systems in the promotion of ethical and trustworthy AI. We will examine the significance of transparency for stakeholders (i.e., users, developers, and policymakers), and the trade-offs between transparency and other objectives, e.g. accuracy and privacy. Recent US legislation prohibiting copyright claims on neural network-generated documents is used to illustrate the issues presented by the opaque nature of black-box AI models.

A thorough literature review is carried out to investigate current approaches and tools for AI transparency and identify gaps and areas for future research.

By moving from black box to glass box AI systems, we can ensure that AI technologies are not only powerful but also ethically sound and aligned with human values. This study adds to the ongoing debate about AI ethics and prepares the way for future research into the complex landscape of transparency, trust, and decision-making in AI systems.

Keywords: transparency · artificial intelligence · black box · neural networks · deep learning ethics · explainable AI · trust · decision-making · algorithmic auditing

1 Introduction

Transparency is increasingly important in AI systems as these technologies permeate various areas, e.g., natural language processing, computer vision, and even AI-generated material (e.g., text documents, images, animations, and music).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14107, pp. 118–130, 2023.
https://doi.org/10.1007/978-3-031-37114-1_9

Transparency is critical in AI systems for fostering trust, promoting fairness, and ensuring accountability, which is critical given their far-reaching effect on society.

Trust is a fundamental aspect of human-AI interactions. As AI systems become more sophisticated and integrated into everyday life, users need to trust these technologies to make informed decisions and engage with them effectively. Transparency in AI systems allows users to understand the underlying decision-making processes, engenders trust and promotes acceptance. E.g., AI-generated content can be used in various applications, from content creation to customer support: ensuring transparency in how these models generate content helps users trust and use their outputs responsibly. Another critical aspect of AI systems is fairness, as biased or discriminatory outcomes can have critical repercussions for people and society [7,8]. Transparent AI systems allow users, developers, and regulators to spot and address potential bias in data or algorithms. Transparency in the context of AI-generated content can help expose any unintended biases in the generated documents, ensuring the fairness of the content given to users. Accountability is another relevant value for AI systems because it allows for the assignment of responsibility for their actions and results. Transparent AI systems enable stakeholders to track decisions back to specific system components, which is critical for holding developers and organizations accountable for any negative outcomes. Transparency in the realm of AI-generated documents can help users, developers, and lawmakers identify the source of any problematic content, allowing for appropriate corrective actions and informed policy decisions.

Unfortunately, transparency in AI systems is not always possible by design because of the black-box nature of the techniques used for training, and the consequent difficulty in providing sources for the final output.

The emergence of many content-generating AI models, e.g. *ChatGPT* [1], *Midjourney* [2], and *Artbreeder* [3], has sparked a revolution in various industries and applications. Initial issues related to accountability are generating scepticism and resistance towards the adoption of AI technologies: if people reject AI systems based on these early challenges, it could result in significant missed opportunities for innovation, efficiency, and societal progress, stifling innovation and impeding the potential benefits that AI can offer. AI-generated content has the potential to transform industries such as journalism, entertainment, education, and marketing, automating tasks and enabling new creative possibilities. AI can indubitably save time and resources, leading to increased productivity and economic growth. AI-generated content can help address global challenges, such as providing access to information and educational resources in underserved areas or creating sample synthetic data sets for domains of research where there is a lack of data (e.g., see Fig. 1 for a transformation of the facial emotion expression of a given image), and where collecting or accessing data may arise ethical issues



Fig. 1. Example of synthetic data transformation of emotional faces made with Artbreeder, [3] starting from a given image (on the left) to a happier one (in the center) and an angry one (on the right), which in this case inherited a cartoonish style from the training data and discrimination model.



Fig. 2. Example of synthetic data of emotional faces created with Midjourney [2].

(e.g., collecting data about unpleasant emotions, [4] which cannot be aroused on purpose, and would be difficult to track, see Fig. 2).

It is critical to understand that the early adoption of AI technologies will unavoidably be fraught with difficulties. These problems, however, ought to be

viewed as chances for development rather than as objections. Governments and stakeholders should put their attention on finding solutions for these problems through research, development, and collaboration rather than ignoring or limiting AI. AI can be used responsibly by developing *glass-box* systems (i.e., supporting additional information to black-box systems, especially for algorithm training), establishing ethical standards, and putting transparency measures in place. This will pave the way for a time when AI-generated content is both accountable and useful. Such a direction could be more beneficial than just avoiding any copyrighted use of AI-generated content.

While the United States government decided that in the USA it is not legit to claim copyright on any document (e.g., text, image, music) generated by AI, [5] ten companies, including *OpenAI*, *TikTok*, *Adobe*, the *BBC*, and *Bumble*, have signed up for a new set of guidelines on building, creating, and sharing AI-generated content responsibly, in the *Partnership on AI* (PAI), an AI research nonprofit [6]. PAI suggestions urge both technology developers and creators and distributors of digitally-created synthetic media to be more open about what the technology can and cannot do, as well as to disclose when people may be interacting with this type of content. The same direction is taken by many researchers, associations, and editors, highlighting such clarity as a relevant outcome of any AI-related research. One of the most essential aspects of the PAI guidelines is the companies' agreement to include research methods to notify users when they are interacting with something generated by AI. A crucial aspect of the guidelines is the commitment of the participating companies to explore and implement methods to inform users when they are engaging with AI-generated content. Potential approaches include the use of watermarks, disclaimers, or traceable elements in the AI model's training data or metadata, highlighting the importance of both transparency and user awareness in the responsible deployment of AI-generated content.

In this paper, the recent development of AI Ethics will be studied providing novel laws and guidelines, and then the state of the art is provided for scientific research on the topic. Some considerations are added to each proposal, and examples are provided of critical issues that remain unsolved. Final conclusions are then drawn.

2 Recent Development of AI Ethics

In recent years, there have been several key developments and milestones in the field of AI ethics. In November 2021, UNESCO's General Conference adopted the Recommendation on the Ethics of Artificial Intelligence, which is the first-ever global standard-setting instrument on the subject [9]. This recommendation aims to protect and promote human rights and human dignity and serves as an ethical guiding compass for building respect for the rule of law in the digital world.

The European Commission suggested a regulation establishing harmonized rules on artificial intelligence (Artificial Intelligence Act) [10] and amending certain Union legislative acts on April 21, 2021. The proposal seeks to guarantee

that AI systems placed on and used in the Union market are safe and adhere to existing legislation on fundamental rights and Union values. It also aims to improve governance and effective enforcement of previously existing laws on fundamental rights and safety requirements applicable to AI systems and to facilitate the development of a single market for lawful, safe, and trustworthy AI applications while preventing market fragmentation. The final rules to be given are still unknown.

Such a plan establishes harmonized rules for the development, market placement, and use of AI systems in the European Union, based on a risk-based strategy. It suggests a single, future-proof definition of artificial intelligence (still not agreed upon), where certain particularly harmful AI practices are prohibited as violating Union values, while specific restrictions and safeguards are proposed in connection to certain law enforcement uses of remote biometric identification systems. The plan establishes a solid risk methodology for defining high-risk AI systems that endanger people's health and safety or fundamental rights. Before these AI systems can be introduced on the Union's market, they must comply with a set of mandatory requirements for trustworthy AI, and conformity assessment processes. No rules are given for researchers, yet.

The ongoing efforts by various organizations and initiatives to establish ethical guidelines and best practices for AI development and deployment, engaging in dialogue with various stakeholders [6], will ideally develop practical solutions for ensuring that AI technologies benefit people and society.

The U.S. Copyright Office has launched an initiative to examine the copyright law and policy issues raised by artificial intelligence (AI) technology [5]. This includes the scope of copyright in works generated using AI tools and the use of copyrighted materials in AI training. The Office convened public listening sessions in the first half of 2023 to gather information about current technologies and their impact and published a notice of inquiry in the Federal Register. The current response is that no copyright can be claimed on AI-generated content.

On February 28, 2020, also the Vatican issued the *Rome Call for AI Ethics*, [15] a document agreed upon by Catholics and leaders of other religions, together with experts from *IBM*, *Microsoft*, and other scientists, outlining six main principles that they all agree should be followed to ensure that AI does not damage humanity, i.e.: Transparency, Inclusion, Accountability, Impartiality, Reliability, Security (and Privacy). The idea is to foster a feeling of shared responsibility among international organizations, governments, institutions, and the private sector to establish a future in which digital innovation and technological progress give humanity primacy.

3 Literature Review of Ethics in AI

The scientific literature on AI ethics has identified several key principles and challenges related to the development and deployment of AI systems. Transparency, privacy, accountability, and fairness are among the most commonly

cited principles, while lack of ethical knowledge and vague principles are significant challenges for considering ethics in AI.

A systematic literature review of AI ethics principles and challenges conducted by Khan et al. [11] revealed that the global convergence set consists of 22 ethical principles and 15 challenges. A lack of ethical knowledge and vague principles were reported as significant challenges for considering ethics in AI.

The study also found that there is still debate about the implications of these principles and that various research organizations, lawyers, think-tankers, and regulatory bodies are involved in developing AI ethics guidelines and principles. The findings of this study are preliminary inputs for proposing a maturity model that assesses the ethical capabilities of AI systems and provides best practices for further improvements.

However, recent studies have found that the current guidelines created for AI ethics are ineffective and are not being implemented in practice, with no impact on ethical decisions [12,13]. There are no tools, methods, or frameworks that bridge the gap between AI concepts and their practical application. More research in this area is needed to discuss AI ethics principles, and challenges, providing evaluation standards and models that guide the AI industry in considering ethics in practice.

Thilo Hagendorff's *The Ethics of AI Ethics: An Evaluation of Guidelines* [14] evaluates 22 key AI ethics guidelines and makes recommendations on how to overcome their relative ineffectiveness. While AI ethics guidelines have been created in recent years to gather principles that technology developers should adhere to, the author contends that these guidelines frequently lack methods to reinforce their own normative claims. The article contends that researchers, politicians, consultants, managers, and activists must address this fundamental flaw in ethics and offers suggestions for improving the efficacy of AI ethics guidelines. Researchers in the field of AI are not considered in the list of relevant subjects for this sensitization idea.

Madhulika Srikumar et al.'s article *Advancing ethics review practices in AI research*, published in Nature Machine Intelligence in 2022, [16] discusses the adoption of ethics review procedures as an important first step in anticipating and mitigating the potential harms of AI research. The authors contend that long-term success will necessitate a coordinated community effort to support experimentation with various ethics review procedures, study their impact, and provide opportunities for diverse voices from the community to share insights and foster norms.

4 Discussion and Considerations on Unsolved Issues

4.1 General Discussion

In the scientific works and the recent contributions from various governments, considered in the previous sections, governments are almost always considered the main guarantors for ethics in AI, but the shortcoming of this point of view is trivially clear: governments can only legislate after they are aware of existing

problems on AI, whereas only researchers in public or private universities and private or supranational research organisations have a real grasp of the problems before they already have the first side effects on society.

Among the others, the Vatican Rome call actually hopes for the emergence of a completely new discipline, i.e. *algorethics*, which is defined as the ethics of algorithms shared by all actors and contains both scientific knowledge and ethical principles as prerequisites for the new knowledge required to ensure AI ethics. This goal seems a long way off: good for raising awareness, but with no immediate practical application.

Scientific research in this matter, instead, provides several works on specific issues, e.g., gender bias, [17] dataset bias [8], overconfidence in AI systems, [7]. Researchers are thus called now, more than ever, at speaking, explaining, and publishing their contribution to the better ethical management of Artificial Intelligence. A critical effort is nowadays invested by researchers in the field of AI to support neural networks and other black-box systems to provide additional information [18] to support the output with sources to guarantee accountability [19] or make the process of data elaboration and classification transparent [20] and explainable, with the recent creation of a new branch of AI called *explanable AI* [21].

4.2 Unsolved Issues and Considerations

The issues that remain unsolved and untreated in the public debate for practical solutions are numerous. The current high hype of ChatGPT, which is arising in acknowledgement of how generative neural networks work and are used to generate content for the final user, highlighted the most evident issues. E.g., there is an objective gap in the law about how to treat the copyright (used in the USA to protect creative products) or author's rights (as it is called in Europe) for content generated by Artificial Intelligence. The US government decided that the personal contribution of the author to write the prompt instructions for the content generation is not enough to state the copyright, because this law is meant to protect personal creations and not ones made by machines. Other countries don't agree and are discussing finding a good compromise, which looks difficult because of the borderline situation, where deciding how much personal creative contribution is generally needed to use a command prompt to generate consistent creations using AI.

A similar problem for copyright can be found by looking at the problem of content generation from a bottom-up point of view: if the content is generated using an algorithm which is trained on material covered by copyright, does the generated document infringe such copyright? The answer could not be trivial: in fact, the main gap is the lack of knowledge about the training data (see the example in Fig. 3 and actual seed image fed as a starting point, used for the generation of a particular document, which is not possible to retrieve for standard black-box neural networks.



Fig. 3. Example of image generated with Midjourney [2] in the style of Van Gogh. Can the application by AI of the style of an existing author infringe the author's copyright? What if a manually painted image is fed to the algorithm as a starting point, will it be as if the image is modified with an editing software?

Furthermore, the hype of generative neural networks is rising interest in almost each and every discipline and field of application: when everybody speaks about generative AI, most of them will speak without the necessary background knowledge, providing the public with countless imprecisions and mistakes that spread virally in the interpretation of the problem, of the right way to use the AI tools, of the goals and skills of generative AI, and, last but not least, of the dangers of AI, which is often seen as the production of a singularity that will self-train, rule the worlds and destroy humanity. Such surreal concerns stemming from an imperfect and insufficient knowledge of the issue are generating

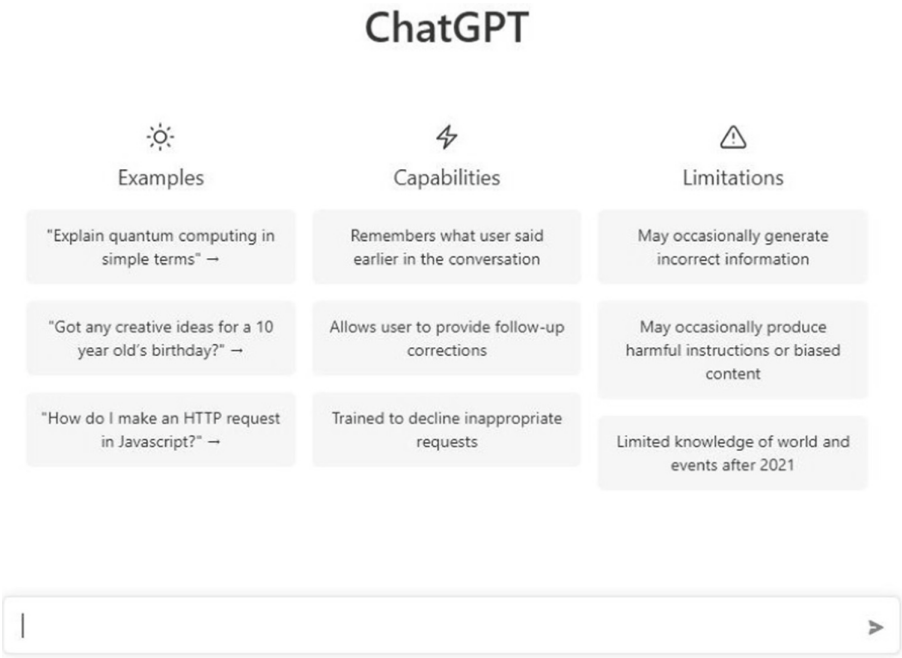

Fig. 4. Homepage of ChatGPT [1]. On the right side the warning about the AI limitations is visible, listing incorrect information, harmful instructions, biased content, and limited knowledge among the potential output results.

serious misunderstandings in the general public and in everyday discussions in the media, which tend to exaggerate in stirring up clickbait controversy. This resonance effect could lead to disdain and rejection by many of the artificial intelligence tools that are being studied to improve the lives of mankind, e.g. in the fields of medicine and wellness. Already today, there is much resistance in the public to believing that what appears to be made to replace human intelligence rather than, as is more correct, to support decision-making or improve the productivity of jobs that are humanly difficult to sustain over a long period of time because they are mentally or physically exhausting. This misunderstanding is producing a rejection similar to that of the AI dark ages of the 1980s, but at a time when the development of AI cannot be slowed down without extreme side effects in research and development.

The need to regulate, rather than AI, the communication and dissemination of information on AI is therefore increasingly evident, to convey correct information in a way that is useful to avoid misuse or inconsistent fears with uncontrolled consequences. A clear example of the need of a different approach to communication is the fact that people do not read the rules and articles of agreement to which they are asked to consent, nut just automatically click the consent button with no clue of its consequences, and then complain about alleged rights they

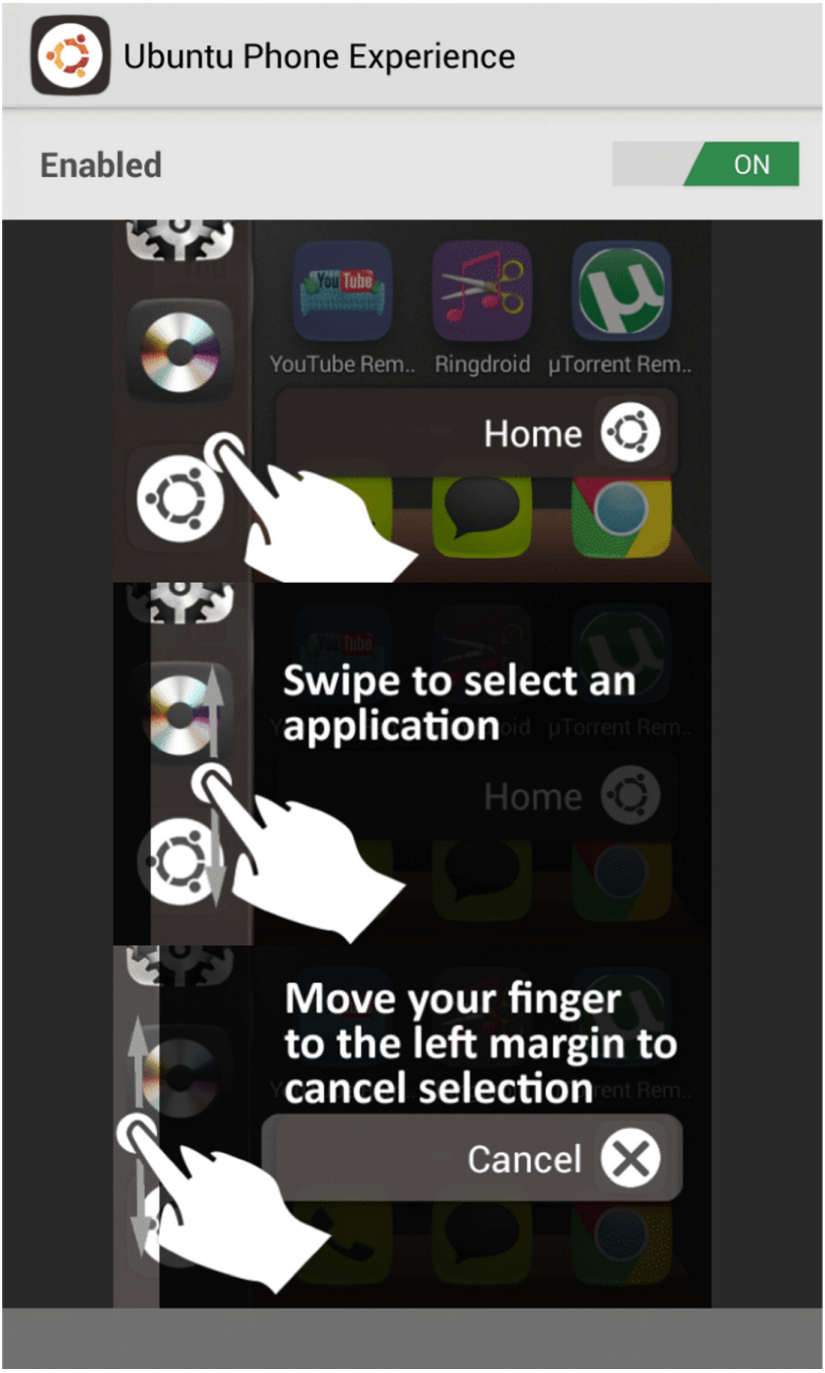

Fig. 5. UI design of a tutorial (in this case, for the Ubuntu phone experience) for the first-time use of an app.

thought were guaranteed and find they are not. Even for the very clear homepage of the online service for ChatGPT (see Fig. 4), where the warning that the system may produce inaccurate, irrelevant, or even dangerous answers is highly visible but ignored by a large number of users, as is clear from their comments on social networks.

It is now more necessary than ever to establish guidelines on the transparent communication of these limitations so that the user is not only alerted but aware of them. One modality could be taken from the tutorials often used for user-interface design when first opening an interface (see the example in Fig. 5), constraining the user to interact with the information provided to him by checking its comprehension (e.g., with clicks, answering basic questions, or gamification).

Concerning misuse, then, a distinction must be made. On the one hand, there is misuse due to insufficient training, for which developers and companies operating AI tools have declared themselves frightened and for which they are already fighting back with guidelines, training, debates and scientific dissemination. On the other hand, there is deliberate misuse, which is already forbidden by law and is not the responsibility of the developers, except perhaps in a tiny part. We can compare this problem to that of developing a bottle: the bottle is designed to contain a liquid and not to be (or not to be) used as a blunt weapon. If there are contexts and situations where it is necessary to regulate to prevent misuse, e.g. by preventing people from entering the stadium with glass bottles or corked water bottles, it is certainly not the duty of bottle designers to put a disclaimer on the label announcing not to throw the bottle at people. In fact, this use is so clearly conscious of illegality that no responsibility can be placed on the designer of the object. The same can be said about AI ethics. On the one hand, it is certainly necessary to train users to correct and conscious use, but surely it must be considered apart from any deliberately criminal use that should be made of generative AI tools, e.g. to generate the animation of a face and make him speak words that he did not say pretending that it is an original video. Such criminal uses are the responsibility of the user and not of the researcher or designer and it is unlikely to expect that a tool will not be made available to the public if it does not prevent any misuse of this kind.

5 Conclusions

The current hype of generative AI raised concerns about Artificial Intelligence that could lead to rejecting AI. There are evident gaps in laws and rules for several issues, e.g., providing sources for content created using generative AIs, copyright, fair use, and education or dissemination of information, among others. Research is the only stage of development where it is possible to intervene before the side effects of AI on society spread.

It is paramount that a debate about ethics in AI is raised among researchers, to avoid misuse and misinformation that could lead to the rejection of AI by the general public opinion. It is important to distinguish misuse due to a lack of

education and information from the voluntary and conscious criminal use of AI tools and provide different solutions to different cases.

Among the issues raised by the recent hype of generative AI, there is the need for a glass-box approach, considering explainable AI to support data for checking the sources, to avoid breach of copyright, and to guarantee the security and privacy of information protected by law.

6 Ethical Statements

AI has been used to support this paper in image generation (i.e., Midjourney), information retrieval (i.e., Bing), language correctness, clarity (i.e., Grammarly), and text translation (i.e., DeepL). No textual content has been included which is generated by AI and not creatively and professionally elaborated by the author's human mind. Figure 5 is under IPRight policy and its use is free, provided that it is clear that this paper is in no way speaking for or on behalf of Canonical and that there is no endorsement by Canonical for this paper. The author acknowledges support by PRIN "PHRAME" Grant n. 20178XXKFY.

References

1. https://openai.com/blog/chatgpt
2. https://www.midjourney.com/
3. https://www.artbreeder.com/
4. Franzoni, V., Biondi, G., Milani, A.: A web-based system for emotion vector extraction. In: Gervasi, O., et al. (eds.) ICCSA 2017. LNCS, vol. 10406, pp. 653–668. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62398-6_46
5. https://copyright.gov/ai/
6. https://syntheticmedia.partnershiponai.org/
7. Franzoni, V., Vallverdù, J., Milani, A.: Errors, biases and overconfidence in artificial emotional modeling. In IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume, pp. 86–90. (2019, October)
8. Franzoni, V., Biondi, G., Milani, A.: Defining classification ambiguity to discover a potential bias applied to emotion recognition data sets. In 2022 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE (2022 December)
9. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
10. https://eur-lex.europa.eu/legal-content
11. Khan, A. A, et al,: Ethics of AI: A systematic literature review of principles and challenges. arXiv preprint arXiv:2109.07906 (2021)
12. Vakkuri, V., Kemell, K.-K.: Implementing AI ethics in practice: an empirical evaluation of the RESOLVEDD strategy. In: Hyrynsalmi, S., Suoranta, M., Nguyen-Duc, A., Tyrväinen, P., Abrahamsson, P. (eds.) ICSOB 2019. LNBIP, vol. 370, pp. 260–275. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33742-1_21
13. Ray Eitel-Porte. Beyond the promise: implementing ethical AI. AI Ethics **1**(1) 73–80 (2021)
14. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. **30**, 99–120 (2020)

15. https://www.romecall.org/the-call/
16. Srikumar, M., et al.: Advancing ethics review practices in AI research. Nat. Mach. Intell. **4**, 1061–1064 (2022)
17. Franzoni, V.: Gender differences and bias in artificial intelligence. In: Vallverdú, J. (ed.) Gender in AI and Robotics. Intelligent Systems Reference Library, vol. 235. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-21606-0_2
18. Räukur, T., Ho, A., Casper, S., Hadfield-Menell, D.: Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. arXiv preprint arXiv:2207.13243 (2022)
19. Kim, B., Park, J., Suh, J.: Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information. Decis. Supp. Syst. **134**, 113302 (2020)
20. Wang, J., Liu, H., Wang, X., Jing, L.: Interpretable image recognition by constructing transparent embedding space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 895–904 (2021)
21. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: a brief survey on history, research areas, approaches and challenges. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 563–574. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_51