

# Ethics of AI

Alessio Tartaro – University of Sassari, a.tartaro@phd.uniss.it

Mihály Héder – Budapest University of Technology and Economics,  
mihaly.heder@filozofia.bme.hu

*This entry presents a comprehensive overview of AI ethics, delving into various ethical and social issues arising from AI's expanding role in different facets of private and societal life. It sheds light on critical challenges such as accountability gaps, biases, discrimination, and risks associated with generative AI technologies which are relevant for information flows and the public communication system. The entry then explores the emergence of concepts like Trustworthy AI and Responsible AI as responses to these challenges, underscoring their importance in developing AI systems that minimize risks and maximize societal benefits. Despite the growing optimism around Trustworthy AI, the entry also emphasizes the difficulties in implementing ethical principles within AI systems. This opens the discussion to critical perspectives that, while not mainstream in AI ethics, offer valuable insights into power dynamics inherent in AI technologies. Furthermore, the article addresses the ongoing debate on AI regulation and standardization, framing it as an extension of the initial AI ethics discourse. Lastly, it outlines the relevance of analyzing public and political discourse on AI to scholars and practitioners in the field of political communication.*

*AI ethics; AI regulation; Trustworthy AI; Social Impacts of AI*

Artificial intelligence (AI) refers to a wide range of computational systems capable of performing tasks - such as understanding human language, playing games, identifying patterns, making decisions etc. - that typically require human intelligence. The ethics of artificial intelligence, or AI ethics, is an interdisciplinary domain that explores a wide spectrum of topics, from philosophical investigations into machine intelligence, consciousness, and morality, to practical impacts of AI on human values such as autonomy, fairness and equal opportunity. Central to AI ethics is also the policy discourse on the regulation of AI.

Given the broad and varied nature of AI ethics, this entry does not claim to be exhaustive. Rather, its goal is to offer readers a well-rounded introduction to AI ethics and its relevance for the field of political communication. To do so, the entry begins with an overview of the key ethical and societal issues related to AI and relevant for political communication. Following this, it delves into the main trends in AI ethics, including its policy implications. Finally, the entry provides some insights on the public discourse on AI, framing it as an interesting subject for political communication and a promising avenue for further research.

### **Ethical and societal concerns in AI**

The discourse on AI ethics predominantly revolves around the multifaceted risks and concerns associated with AI technologies. Some of these risks are considered pervasive. Accountability gaps can arise in any context where decisions are automated. The opacity of AI systems prevents a full understanding of how an AI system arrives at its decisions or recommendations. Related to this is the risk that AI may lead to the perpetuation of bias and discrimination due to the ingrained prejudices that are often embedded in the training data. Regardless of the specific AI applications and use cases, these risks are paramount and central to the discussion of AI ethics. However, some more specific risks can be identified that are relevant in the field of political communication.

The emergence of deepfakes - i.e., fake yet realistic texts, images, and videos created by generative AI - exemplifies this novel kind of risk. When utilized in political contexts, deepfakes raise serious concerns regarding the spread of misinformation and the manipulation of public perception (Hancock et al., 2021). Research indicates that when faced with deepfake videos of political figures, individuals struggle to discern their authenticity (Hameleers et al., 2024). Intriguingly, when participants do correctly identify a deepfake, their success is not attributable to characteristics of the manipulated audiovisual content.

Instead, they recognize inconsistencies between the arguments presented in the political deepfake and the established profile of the depicted politician. This suggests that a solution to the risks posed by deepfakes may not reside in technical tools for their detection, but rather in the development of effective media literacy interventions.

Such interventions are even more crucial as they could aid in preserving trust in information flows and the public communication system. Indeed, the most significant risk in the dissemination of deepfakes is the escalating skepticism towards the reliability of online information. Over time, the pervasive expectation that much of the online content could be false may diminish cooperative and responsible behavior. This, in turn, threatens to erode the foundations of meaningful public discourse (Vaccari & Chadwick, 2020). It is in response to these challenges that AI ethics endeavors to foster the development of AI resilient to abuse and misuse.

### **Trustworthy AI, ethics guidelines and AI principles**

Concepts like Trustworthy AI have emerged in the scholarly and policy discourse to refer to AI systems that maximize benefits and minimize risks for individuals and societies. This notion has been championed by governments, international organizations, standard-setting bodies, and industry players alike. A key contribution to this discourse are the “Ethics Guidelines for Trustworthy AI”, produced by a group of independent experts appointed by the European Commission. Hundreds of similar documents have been published by various stakeholders, outlining ethical principles that AI systems should follow.

The structure and objectives of these guidelines share remarkable similarities and highlight notable communication patterns. They typically begin with the assertion, often taken for granted but rarely substantiated, that AI brings substantial benefits, positioning it as a revolutionary technology that societies should gradually accept and adopt. Concurrently, they caution against a wide range of risks, some of which were reviewed in the previous section of this entry. Finally, these guidelines proceed to outline a series of principles that AI systems should adhere to in order to be Trustworthy. They include, for example, principles such as transparency, explainability, fairness, accountability, and privacy (Jobin et al., 2019).

Despite the growing number of guidelines, consensus on the efficacy of these guiding principles is not universal. A growing body of critical perspectives is beginning to challenge the prevailing views and offers alternative approaches.

## **Principle-implementation gaps and other critical approaches**

The guidelines have the merit of raising a discussion about the need to include ethical, societal and fundamental rights considerations in AI research, development and use, but they also have clear limitations. In the field of AI ethics, a growing number of critics argue that the current trend merely provides broad, high-level principles without offering any solutions on how to apply these ethical principles in concrete situations. According to Mittelstadt (2019), principles alone cannot guarantee ethical AI because of the lack of methods to translate these principles into practice. More radically, Hagendorff (2020) argues that the operationalization of these principles by technical means would require a simplification of the multidimensionality of these complex ethical concepts to something that is measurable, and calculable. Finally, Héder (2020) argues that, except for transparency and human oversight, most of the principles are not AI-specific, i.e., they would apply to many other technologies, and they are too abstract to be effective. These critical findings raise the need to bridge the gap between principles and practice.

Other scholars argue that this challenge is inherent to the mainstream approaches in AI ethics and cannot be effectively addressed by merely devising technical implementations of these principles (Munn, 2023). From this viewpoint, the core issue is not the absence of technical operationalization of ethical principles, but rather the fundamental discordance between ethical aspirations and the capitalist mode of production that underpins AI development. The development of AI should be interpreted as a manifestation of power dynamics, where private entities and governments compete for control over this potent technology. This perspective advocates for a shift towards approaches that emphasize the material and political dimensions of AI, thereby reconceptualizing AI as an extractivist technology (Crawford, 2021).

## **Policy implications**

The rapid evolution of AI, coupled with growing awareness of its risks as discussed in the field of AI ethics, has underscored the necessity for regulatory measures. Major players in the AI arena have adopted regulations for these technologies. For instance, the US promulgated an “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence” in October 2023. In January 2023, “Administrative Provisions on the

Management of Deep Synthesis of Internet Information Services” to regulate deepfakes entered into force in China. In March 2023, the European Union adopted the AI Act. The latter Regulation has garnered significant attention. The AI Act introduces a risk-based regulatory approach, imposing progressively stringent requirements on AI systems relative to their level of risk. AI applications deemed to pose unacceptable risks, such as those used for social scoring by governments, are outright prohibited. For high-risk AI systems, like those employed in healthcare, labor, and education sectors, the AI Act stipulates comprehensive obligations concerning risk management, data quality, human oversight, and transparency, among others.

In addition to formal regulations, the conversation around AI policy also includes forms of industry self-regulation and co-regulation. Within this discourse, the topic of AI technical standards is gaining prominence. Standards are voluntary technical specifications that industry stakeholders can adopt to ensure interoperability, maintain product quality, and manage risks. Under the AI Act, technical standards are expected to play a crucial role to implement the Regulation's requirements (Tartaro, 2023). Moreover, given the ethical and societal implications of AI systems, development of standards in this domain is also advancing, with initiatives like the IEEE 7000 series proposing methodologies for integrating ethical considerations into AI system design.

### **The public discourse on AI as a subject of political communication**

For scholars and practitioners in the field of political communication, the construction of political discourse surrounding AI presents a fascinating area of study. The way governments and public figures depict “AI futures” reveals much about the narratives prevalent in public discourse and the envisioned power dynamics in future AI-influenced societies (Köstler and Ossewaarde, 2022).

National AI strategies are valuable resources for analyzing how different countries frame the discourse on AI. In Europe, for example, these strategic documents often portray AI as a quasi-mythological benevolent force that can help nations in achieving their political objectives. European countries often depict AI as a panacea to national challenges, thereby promoting techno-solutionist narratives that obscures potential adverse aspects of AI, such as environmental degradation, digital surveillance, concentration of power, and the erosion of democratic processes (Ossewaarde & Gulenc, 2020). Similar narratives have been observed in the South Korea’s AI national strategy as well (Kim, 2023).

An examination of AI policy documents on an international level, covering 37 countries across all continents, reveal a more diverse set of governmental narratives. This analysis demonstrates that different stories are told about AI: each country crafts its unique story around AI, characterized by unique settings, a cast of diverse characters, intricately woven plots, and the assignment of varied roles to governments, all aimed at teaching distinct lessons or morals (Guenduez & Mettler, 2023).

Policy narratives and communications are important as they strategically build a shared understanding of matters of public significance. In this regard, it is significant to observe how political leaders in Europe have echoed the narrative concerning the “existential risks” of AI, emphasized by some industry leaders such as Elon Musk, founder of xAI, and Sam Altman, CEO of OpenAI. For instance, von der Leyen, 13th president of the European Commission, highlighted existential risks from AI in her 2023 State of the Union address, and these concerns were central to discussions at the 2023 AI Safety Summit sponsored by the UK government. The focus on existential risks, however, is often perceived as a potential deflection from more immediate and tangible AI-related concerns, such as discrimination and lack of accountability, issues that are increasingly manifesting. Consequently, some scholars have posited that the current trends in AI ethics and regulation may represent a form of regulatory capture (Saltelli et al., 2022), rather than a balanced response to the diverse array of risks posed by AI.

## **Conclusions**

This entry has provided a comprehensive overview of the principal ethical issues associated with AI, and how this intersects with the field of political communication. Given the ongoing advancements in technology, it is inevitable that these issues will evolve, with new challenges emerging over time.

AI as a subject of political communication, in particular, presents an interesting future avenue for research. Existing studies have already yielded noteworthy insights through the examination of national AI strategies. This analytical approach could be broadened to encompass other forms of communication utilized by policy actors.

Furthermore, the way the media engage with the discourse on AI constitutes another developing field of study, holding the promise of revealing significant developments (Nguyen & Hekman, 2022). Lastly, the potential for marginalized and minority groups, often adversely

impacted by the negative consequences of AI, to forge and disseminate a counter-narrative to prevailing AI narratives represents an area that remains largely unexplored.

## References

- Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. <https://doi.org/10.12987/9780300252392>
- Guenduez, A. A., & Mettler, T. (2023). Strategically constructed narratives on artificial intelligence: What stories are told in governmental artificial intelligence policies? *Government Information Quarterly*, 40(1), 101719. <https://doi.org/10.1016/j.giq.2022.101719>
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, 30(1), 99-120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hameleers, M., Van Der Meer, T. G. L. A., & Dobber, T. (2024). They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes. *European Journal of Communication*, 39(1), 56–70. <https://doi.org/10.1177/02673231231184703>
- Hancock, J. T., & Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152. <https://doi.org/10.1089/cyber.2021.29208.jth>
- Héder, M. (2020). A criticism of AI ethics guidelines. *Információs Társadalom: Társadalomtudományi Folyóirat* 20 (4), 57-73. <https://dx.doi.org/10.22503/inftars.XX.2020.4.5>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kim, J. (2023). Traveling AI-essentialism and national AI strategies: A comparison between South Korea and France. *Review of Policy Research*, 40(5), 705–728. <https://doi.org/10.1111/ropr.12552>
- Köstler, L., & Ossewaarde, R. (2022). The making of AI society: AI futures frames in German political and media discourses. *AI & SOCIETY*, 37(1), 249–263. <https://doi.org/10.1007/s00146-021-01161-9>
- Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1, 501–507 (2019). <https://doi.org/10.1038/s42256-019-0114-4>
- Munn, L. The uselessness of AI ethics. *AI Ethics* 3, 869–877 (2023). <https://doi.org/10.1007/s43681-022-00209-w>

- Nguyen, D., & Hekman, E. (2022). The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation. *AI & SOCIETY*.  
<https://doi.org/10.1007/s00146-022-01511-1>
- Ossewaarde, M., & Gulenc, E. (2020). National Varieties of Artificial Intelligence Discourses: Myth, Utopianism, and Solutionism in West European Policy Expectations. *Computer*, 53(11), 53–61. <https://doi.org/10.1109/MC.2020.2992290>
- Saltelli, A., Dankel, D. J., Di Fiore, M., Holland, N., & Pigeon, M. (2022). Science, the endless frontier of regulatory capture. *Futures*, 135, 102860.  
<https://doi.org/10.1016/j.futures.2021.102860>
- Tartaro, A. (2023). Regulating by standards: Current progress and main challenges in the standardisation of Artificial Intelligence in support of the AI Act. *European Journal of Privacy Law & Technologies*, 1, Article 1. <https://doi.org/10.57230/ejplt222AT>
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1), 2056305120903408. <https://doi.org/10.1177/2056305120903408>