American Philosophical Quarterly
Volume 61, Number 4, October 2024

INTRODUCTION TO THE SPECIAL ISSUES

The Ethics of Artificial Intelligence: Exacerbated Problems, Renewed Problems, Unprecedented Problems

Luciano Floridi

abstract Artificial intelligence (AI) is rapidly reshaping our world. As AI systems become increasingly autonomous and integrated into various sectors, fundamental ethical issues such as accountability, transparency, bias, and privacy are exacerbated or morph into new forms. This introduction provides an overview of the current ethical landscape of AI. It explores the pressing need to address biases in AI systems, protect individual privacy, ensure transparency and accountability, and manage the broader societal impacts of AI on labor markets, education, and social interactions. It also highlights the global nature of AI's challenges, such as its environmental impact and security risks, stressing the importance of international collaboration and culturally sensitive ethical guidelines. It then outlines three unprecedented challenges AI poses to copyright and intellectual property rights; individual autonomy through AI's "hypersuasion"; and our understanding of authenticity, originality, and creativity through the transformative impact of AI-generated content. The conclusion emphasizes the importance of ongoing critical vigilance, imaginative conceptual design, and collaborative efforts between diverse stakeholders to deal with the ethical complexities of AI and shape a sustainable and socially preferable future. It underscores the crucial role of philosophy in identifying and analyzing the most significant problems and designing convincing and feasible solutions, calling for a new, engaged, and constructive approach to philosophical inquiry in the digital age.

keywords Artificial Intelligence, ethics, technological innovation, social impact, sustainability.

1. Introduction

Let me start with some obvious remarks, the kind of remarks we are getting used to receiving by any decent chatbot. I shall be brief because they are platitudes.

In the past few years (ChatGPT 3 was made available in March 2022), artificial intelligence (AI), almost often equated to machine learning, has emerged as a leading and disruptive innovation, ushering in a new era of capabilities and transformations. Its staggering growth in power and applicability heralds the dawn of an era marked by computational systems that have the potential to learn, adapt, and work (e.g., generate content, make decisions, and control other systems) autonomously. AI's profound impact is felt across all sectors of human endeavor—from revolutionizing healthcare diagnostics to streamlining financial systems—reshaping industries, economies, jobs, education, conflicts, cultures, politics, and the very fabric of societies globally. All this exacerbates old ethical problems, reshapes some of them, and creates new ones. As AI systems gain increased autonomy and become integral to all kinds of processes, fundamental ethical issues such as accountability, transparency, bias, and privacy expand their profound, negative impact and widen their scope, morphing into new societal challenges. They demand a conceptually sophisticated understanding of the confluence of technology, ethics, and innovation. This is the first of three special issues of the *American Philosophical Quarterly* intended to address such a challenge.

I will not summarize the articles' contents or try to tease out common themes. The first exercise is pointless (that is what abstracts are for), and the second is too tough, once you grasp the breadth of topics covered and the multidisciplinary nature of the approaches and methodologies used to address them. Of course, I could have asked ChatGPT to do it, but so can the reader. Instead, what may help to navigate such an impressive wealth of studies and reflections is some background that can contribute to contextualizing them in the current debate, and some highlights, to see where the debate may be heading.[1]

2. Exacerbated and Renewed Problems

Let me start from the points already emphasized above. When AI systems implement decision-making autonomy, abilities to learn and improve their performance, and can interact with humans with increasing sophistication, ethical considerations concerning accountability and transparency become pressing. Who is responsible for an autonomous AI system making a significant error, including blatant hallucinations? How do we ensure that AI systems are transparent and accountable, upholding ethical standards and safeguarding against harm? Indeed, are they even explainable in the first place?

Any reader of this introduction knows that a major ethical challenge in AI's development and deployment is the prevalence of bias. AI systems, trained on vast datasets, often reflect and amplify societal biases inherent in their training data. Call it the BIBO problem: bias in, bias out, to rephrase a well-known saying (GIGO: garbage in, garbage out, which refers to data quality). Such biased algorithms lead to discriminatory outcomes, further entrenching societal disparities and injustices. For instance, facial recognition technology has been repeatedly criticized for failing to recognize individuals correctly and unbiasedly. This is not just a technical issue but a profound ethical failure that perpetuates existing societal prejudices. To counteract bias in AI, a coordinated effort is crucial to promote fairness, transparency, and accountability across all stages of AI creation, from data gathering and model training to real-world application and ongoing assessment. We must engage in proactive measures, such as diverse and inclusive data collection, quality controls, regular auditing of AI systems for biased outcomes, and the development of algorithms that can detect and correct biases. This also involves diversifying the AI field, ensuring the teams developing these systems represent various perspectives and backgrounds. Of course, as always, it is human decisions that ultimately count.

The ethical implications of AI also spill over into privacy, consent, and data protection concerns. We do not come from a great past: we are so used to social media using our data, for example, and digital surveillance is now a daily experience for millions of people. However, the future looks even more challenging because privacy concerns are amplified in the age of AI due to the technology's ability to aggregate, analyze, and cross-reference data at an unprecedented scale and speed. Personal data, once somewhat more siloed and less accessible to correlate, can now be combined to create detailed profiles of individuals' lives, preferences, choices, tastes, and inclinations, and even anticipate them, raising worries about surveillance and data security, everywhere, but especially in non-democratic and illiberal countries and contexts. As AI systems amass, analyze, and manipulate immense amounts of personal information, questions emerge about individual rights and safeguarding sensitive data, especially in contexts like education, work, and public health. For example, we know that an AI system designed for healthcare should prioritize values such as patient confidentiality, informed consent, and non-maleficence. These values must be embedded into every layer of the system, from how patient data are collected and stored to the algorithms that analyze and interpret those data to make diagnostic recommendations. AI developers and regulators must work together to ensure that privacy is not sacrificed at the altar of convenience (ours—we are all sinners who cannot throw the first stone), efficiency (of the service providers), or control (by the powers to be). This may involve developing privacy-friendly AI systems (the debate on ethics by design is growing), implementing robust data protection measures, and ensuring that individuals, or institutions on their behalf, have control over their data and how AI systems use them, while always keeping in mind that what needs to be protected are the individuals themselves, not their data.

Removing bias and protecting privacy require accountability and transparency, two other cornerstones of the ethics of AI. As AI systems and their decision-making processes become more complex, ensuring that they are transparent and that their operators are accountable becomes increasingly pressing. This is why explainable AI has emerged as a crucial field dedicated to making AI decision-making processes understandable to humans, which is critical for both trust and accountability when things go wrong, and to improve AI systems when we wish to perform differently or better (see "machine unlearning").

The concept of accountability extends beyond understanding AI's decisions to encompass moral responsibility and legal liability for those decisions. In the event of a failure or harm caused by an AI system, determining who is responsible, and to what degree, may be complicated. Traditional models of responsibility may not suffice, as the "decision-maker" may not be one or more humans in or on the loop, but an AI system left to operate autonomously, potentially influenced by developers, data scientists, users, and even other AI systems, and transformed by its learning process. For example, as the development and deployment of weaponized AI and autonomous weapon systems (AWS) in war zones increases—the conflicts in Ukraine and Gaza offer tragic contexts for experimentation—attributing accountability becomes more challenging but also essential, especially when those actions lead to unpredictable outcomes or result in unintended or unjustifiable harm. We may need to develop more distributed models of responsibility, to account for the multiple agents involved in the design, development, deployment, and operation of AI systems and address satisfactorily the so-called "responsibility gap" within the bounds of international humanitarian law and ethical norms. Likewise, we may need new frameworks for moral responsibility and legal liability that reflect the distributed nature of AI decision-making. For example, the future looks open to conceptual innovation about what we mean by strict liability.

So far, I have outlined some old ethical problems that have been made worse or reshaped by the AI revolution. However, any ethical overview of AI would be incomplete without considering its broader societal impacts. Like other technologies in the past, AI systems are taking on tasks traditionally performed by humans, so they have the potential to reshape drastically labor markets, education systems, and social interactions. The difference with previous revolutions, like the agricultural and the industrial ones, is that those took millennia and centuries (respectively) to impact societies fully. In contrast, the digital revolution, including AI, is transforming the world in the span of a single generation—ours. The resulting economic and social changes are not only profound but also incredibly fast-paced. We must anticipate and manage them now, to prevent exacerbating inequalities and social fractures. We cannot afford to address them *a posteriori*, when the human, environmental and financial costs to fix any negative impact will be astronomical. Brown collars (agriculture, or "bioware") and blue collars (industry, or "hardware") were replaced by the arrival of the engine. White collars (services, or "software")—representing more than 90 percent of the workforce in the US, for example—seemed safe and poised to grow. However, the AI revolution is automating white-collar jobs and can potentially displace workers, leading to unemployment and economic hardship. This is an unemployment not due to lack of demand—there are plenty of jobs in any advanced information society with a strong AI presence, like the US, South Korea, or Singapore—but rather due to a mismatch between demand and supply (people lacking the right skills or expertise) and a widening of the so-called "digital divide," in this case represented by the inequality in education and access to AI services. Future jobs will be in the management of digital technologies by green collars. However, to avoid or mitigate adverse effects, policymakers and industry leaders must consider strategies such as educational reforms, retraining programs, and other social safety nets for the generation that will bear the brunt of such a rapid revolution. In other words, some of AI's future financial advantages must be spread to tackle the problems we face now.

AI's impact is not only amplifying old digital problems and creating new societal challenges, but also democratizing risks (in the computer science sense of the world: everybody is impacted) and globalizing issues. After all, the ethics of AI cannot be confined to a single cultural or national perspective. Consider the following two examples.

The environmental impact of AI concerns the sustainability of AI technologies, the energy consumption and carbon footprint of data centers, the production and operation of AI systems, and the broader digital infrastructure underpinning these technologies. Data centers are the backbone of AI and demand massive amounts of electrical power for operational purposes and cooling systems, thus contributing to greenhouse gas emissions. The production of AI hardware and devices incorporates resource-intensive processes, including the extraction of rare earth minerals and water consumption, leading to the degradation of ecosystems and pollution. To make matters worse, the replication and proliferation of AI technologies exacerbate these impacts through a cumulative effect: more devices, more energy consumption, and greater emissions. This raises urgent questions about the sustainability practices within the tech industry and the need for green computing solutions. Improving energy efficiency, using renewable energy sources in data centers, implementing recycling and repurposing strategies for AI systems and devices, and using AI itself to support the UN sustainable development goals, are reasonable strategies. A responsible and ethical approach to AI innovation that harmonizes technological progress with environmental stewardship, to mitigate the ecological footprint of AI, is possible and should be pursued. The alliance between the Green of all our natural and artificial environments and the Blue of all our digital technologies, especially AI, must be the human project for the twenty-first century.

Second example. The advancement of AI technologies is introducing significant security risks by increasing vulnerability to sophisticated cyber-attacks. Malicious actors can exploit AI systems to develop more advanced hacking techniques: AI-powered phishing attacks that are highly personalized and difficult to detect, automated exploitation of software vulnerabilities at scale, and the use of machine learning to breach encryption systems more efficiently. Moreover, AI systems themselves can become targets, with attackers seeking to manipulate AI algorithms through techniques like data poisoning or adversarial attacks, aiming to degrade system performance or induce erroneous outcomes.

In both cases, we are facing global challenges that must be tackled while also dealing with the added complexity of diverse cultural norms and values. We need international collaboration to establish ethical guidelines, technical standards, and legal rules that respect different cultural contexts, which often complement each other, while upholding universal human rights.

So far, I have emphasized exacerbated or renewed problems, but AI is also generating unprecedented problems, some of which are already apparent. Three of them are probably more pressing than others in terms of how they will transform our culture. I shall sketch them out separately, even though they are intertwined.

3. Unprecedented Problems: Copyright and Intellectual Property

AI poses unprecedented challenges to copyright and intellectual property (IP) rights, fundamentally reshaping the landscape in ways that existing ethical and legal frameworks struggle to accommodate. AI's ability to produce works that were previously the domain of human intellect challenges traditional notions of creativity and authorship, leading to significant legal and ethical dilemmas. Three areas are particularly pressing.

First, on the input side, AI complicates the enforcement of IP rights. AI systems are trained on, and can process and learn from, vast datasets, including copyrighted materials, without clear guidelines on the legality of such practices, often risking infringing upon existing copyrights, also because AI can generate content that is substantially similar to works within the datasets used for training. Determining liability in such scenarios—whether it falls on the developers, users, or AI—remains a contentious issue.

Second, in terms of its output, AI generates works—ranging from texts to pictures, from videos to music—that raise the question of whether creations by AI systems alone or just supervised by humans may be protected under copyright laws, which traditionally require a human author. The notion of authorship is thus expanded, pressing lawmakers to rethink the criteria for copyright eligibility. This is compounded by the difficulty in tracking and proving the originality of AI-generated content, given the technology's capacity to generate vast amounts of derivative works easily, rapidly, and cheaply.

Third, AI's role in *deepfakes* (hyper-realistic digital falsifications) presents acute challenges in protecting individual rights and trademarks, damaging reputations, combating misinformation, and misleading the public.

Addressing all these challenges means reconsidering existing IP frameworks to accommodate the realities of AI. This might include adopting more flexible copyright standards, considering new forms of IP rights tailored for AI-generated content, developing international agreements to manage AI-driven cross-border IP issues, using watermark or software solutions to certify authorship and provenance, and upgrading the fair use framework (creative commons) when it comes to content on which AI is trained. These are just examples of ongoing debates. Solutions will have to be embedded within a framework that safeguards human creativity, individual rights, and cultural heritage.

4. Unprecedented Problems: Hypersuasion

By offering unprecedented personalization, efficiency, and predictive power, AI (e.g., recommender systems) is also reshaping social dynamics, communication, and socio-political life, presenting new challenges regarding autonomy erosion, manipulation of freedom of choice, and the power of persuasion. This pervasive ability to limit and shape individuals' capacity to make free and independent decisions, subtly guiding their thoughts, beliefs, emotions, and actions in directions predetermined by underlying algorithms, may be called *hypersuasion*, the hyper-persuasion that AI can exercise on each of us.

AI's hypersuasion can dictate the information, news, and viewpoints to which people are exposed. By determining the content that individuals see and interact with, AI systems can foster *echo chambers* and *filter bubbles* that reinforce existing beliefs and prejudices, polarizing views, and debates, thus stifling constructive dialogue and understanding between disparate groups. AI's hypersuasion can also suggest options that shape preferences, from the trivial choice of the next movie to watch to the future school or job that is best for a person, from the food and lifestyle that may become more appealing to the political orientations that are presented as more convincing. This manipulation extends into the political domain, where AI's hypersuasion can sway public opinion and voting behavior by selectively presenting or withholding information, influence elections through micro-targeted campaigns and deepfakes, and manipulate political narratives and public perception, ultimately undermining the integrity of democratic systems. This fragmentation of the public sphere poses risks to social cohesion and the democratic process, especially when AI is used to disseminate misinformation or polarizing content.

5. Unprecedented Problems: The Future of Content

AI's automatic production of increasingly high-quality content, quickly and cheaply, presents novel and complicated challenges, with epistemological, ethical, and socio-cultural implications that we are only beginning to understand. We saw that AI's content creation raises epistemological questions about authenticity, originality, and creativity. We must reconsider and upgrade these and related concepts in light of the AI revolution. We cannot simply ignore them or pretend that old solutions, as successful as they might have been, only need to be extended to solve new challenges.

From an ethical perspective, consider, for example, that we come from a consumeristic culture that has overemphasized the value of the product or output for almost a century—it does not matter who made a t-shirt, how and why, we only check its quality and price. This culture is now forced to rethink the importance of the process and the source of the product. For example, it is not just the content and style of a diary page that may count, but also, and I would say above all, who wrote it, why and how. Because one may not be able to tell whether the short text is human-generated or AI-generated, or it may soon be indistinguishable, but it matters profoundly whether Anne Frank or ChatGPT wrote it. Finally, we already encountered the societal impact of AI. The automation of content production cannot be underestimated. As these technologies democratize the ability to create high-quality content, they also threaten to disrupt traditional industries and livelihoods.

6. Conclusion

Looking to the future, it is evident that the ethical landscape of AI will continue to evolve as technology advances. In this brief introduction, I only outlined some well-known ethical problems and highlighted a few issues that deserve special attention because their impact will grow significantly in the near future, and shape our cultures. We must remain not only critically vigilant to ensure that our ethical frameworks can accommodate and shape new developments, but also imaginative and innovative in our *conceptual design* (my definition of philosophy) of the best solutions that can offer a good chance to develop a sustainable and socially preferable future. It is all doable. However, analyzing problems and synthesizing their solutions will more likely succeed if we engage with a wide range of stakeholders, from technologists and ethicists to policymakers and the public, to navigate the moral complexities of AI insightfully and tolerantly. This is good news for philosophy: there is plenty of excellent and crucial conceptual work to be done, collaboratively, foundationally, and innovatively. We desperately need a new philosophy of our time for our time, that not only excavates, deconstructs, analyses, unveils or suspects old and new questions, but also produces, builds, constructs, supports and defends new solutions. Philosophy is sorely needed to identify and analyze the most significant challenges posed by the digital revolution and design their convincing and feasible solutions. There is just a final "but." But for this to happen, philosophy must step out of its ivory tower. And on that note, some cautious optimism comes from the authors' contributions to these three special issues.

Yale Digital Ethics Center
Yale University
85 Trumbull Street, New Haven, CT 06511

Department of Legal Studies
University of Bologna
Via Zamboni 22, Bologna, Bo 40100
luciano.floridi@yale.edu

NOTE

[1] The reader interested in knowing my positions about the problems outlined in this introduction may find Floridi (2023a, b) relevant.

REFERENCES

Floridi, Luciano. 2023a. *The Ethics of Artificial Intelligence—Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Floridi, Luciano. 2023b. *The Green and the Blue—Naive Ideas to Improve Politics in an Information Society*. New York: Wiley.