AI regulation and policy-making: ethical and legal issues on unstable ground[a]

Paolo Monti[†], *Norberto Albano*[‡]

1. Putting AI to the Regulatory Test

Calls for the regulation of artificial intelligence (AI) technologies are numerous and have been raised by a wide range of actors, from civil society advocacy groups to political movements and even industry leaders, who, in a surprising move, are simultaneously pushing AI-based services to the public at an unprecedented pace and calling for top-down regulation. The strategic ambivalence of tech corporations can be interpreted in two ways: on the one hand, as an attempt to shape the regulatory field in their own favour pre-emptively (*regulatory capture*), for instance by promoting safety and compliance standards so burdensome in terms of computational and human resources that they create *de facto* barriers to entry for smaller competitors and start-ups; on the other hand, as a maneuver to delegate ultimate responsibility for the negative externalities generated by their own creations to the public sphere. These externalities include not only labour displacement and social polarisation but also the erosion of individual autonomy through personalised persuasion systems, environmental impact, and the degradation of the information ecosystem. Thus, a paradox emerges of a power that, at its peak of expansion, demands to be contained, yet ends up defining the very terms, language, and boundaries of that limitation itself.

The lack of specific legal frameworks and regulations has initially left a lot of room for case-by-case initiatives by regulators and educational institutions, along with a fairly broad academic and public debate about the risks and opportunities. After an early phase of fragmented interventions, more comprehensive regulatory projects have emerged, such as the European Union's Artificial Intelligence Act, which formally entered into force in 2024, the late 2023 White House Executive Order, and the United Nations' report *Governing AI for Humanity*[1]. The urgency of these regulatory decisions is confronted by the vast diversity of AI technologies and applications, but also, and perhaps more fundamentally, by the very unstable foundations on which the normative discourse on the use of AI has to be built, both conceptually and practically. The cornerstones of modern law rest on concepts such as individual autonomy, rationality, and intentionality, attributed to

[a] Published 09/12/2025.
[†] University of Milan-Bicocca, e-mail: paolo.monti@unimib.it
[‡] University of Turin, e-mail: norberto.albano@unito.it
[1] For reference, see: European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, 2024; The White House, *Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 2023; UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021.

identifiable human subjects. AI systems, acting as quasi-autonomous and often inscrutable entities, crack these pillars, creating a dual tension: one of a conceptual nature, which questions the validity of our fundamental legal categories, and one of a practical-temporal nature, which challenges the very ability of the law to keep pace with technological change. This situation creates a political dilemma in which legislators find themselves needing to provide rapid responses to risks perceived as imminent, while lacking the conceptual lexicon and practical tools adequate to regulate such an elusive and metamorphic object.

2. The fragile conceptual foundations

The first of these tensions, as we have mentioned, is *conceptual*: the rapid advancements in AI technologies challenge the applicability of traditional concepts such as responsibility and accountability, authorship and creativity, intelligence and action, personhood and subjectivity, which form the fundamental lexicon of much of the normative discourse. Notions such as responsibility and liability, developed for human agents, undergo a process of dissolution[2] when faced with algorithmic decision chains characterised by opacity, distribution, and emergent properties. The opacity of large *deep learning* models - the so-called "black box problem"[3] - creates an almost insurmountable evidentiary gap: how can an injured party prove the designer's negligence if even the designer cannot reconstruct the process that led the system to its error? In criminal law, the impossibility of attributing a non-human *mens rea* evokes the ancient tensions of the legal maxim *societas delinquere non potest*, here in a new machinic guise.

This tension, defined by some authors as "conceptual disruption"[4], also manifests itself forcefully, for instance, in the domain of intellectual property. Although cases like *Thaler v. Perlmutter* have denied copyright protection to works generated without a human author[5], a more nuanced position emerged in the case of the comic *Zarya of the Dawn*, created using AI, where protection was granted to the selection and arrangement of the images, but not to the individual images generated by Midjourney. This decision introduces the criterion of "sufficient human creative contribution" and notes an excessive distance between the user's input (*prompt*) and the machine's output, likening the AI to a commissioned party rather than a mere tool like a camera. Upstream, the very legality of training these models on protected data is at the heart of landmark legal battles, such as *Authors Guild v. OpenAI*, where the companies' defence is based on the US doctrine of fair use, controversially balancing the "transformative" use of data to create a new model against its market impact on the original works.

---

[2] A. Matthias, The responsibility gap: Ascribing responsibility for the actions of learning automata, in «Ethics and Information Technology», 6, n. 3, 2004, pp. 175-183.

[3] J. Burrell, How the machine "thinks": Understanding opacity in machine learning algorithms, in «Big Data & Society», 3, n. 1, 2016, pp. 1-12.

[4] G. Löhr, *Conceptual disruption and 21st century technologies: A framework*, in «Technology in Society», 74, 2023; S. Marchiori, K. Scharp, *What is conceptual disruption?*, in «Ethics and Information Technology», 26, 2024; J. Hopster et al., Conceptual Disruption and the Ethics of Technology, in S. Nyholm, J. Hopster, P. Lemmens (edited by), *The Ethics of Socially Disruptive Technologies: An Introduction*, Open Book Publishers, Cambridge 2023, pp. 141-162.

[5] United States Court of Appeals for the D.C. Circuit, *Thaler v. Perlmutter*, in «Federal Reporter, 4th Series», vol. 130, 2025, p. 1039.

A structural paradox emerges that strikes at the heart of the relationship between law and technology. In copyright law, a *high degree of human control* over the creative process is required to grant authorship and its ensuing rights. In the attribution of liability, *significant control* over the system's behaviors is demanded to assign blame and the duty of compensation. Since the current technological and commercial advancement of AI aims precisely to increase its autonomy and, consequently, reduce the need for direct human control, the technology itself is moving inexorably toward a grey area, a vacuum of accountability. The foundational distinctions between intelligence and calculation, between action and automation, and between subjectivity and simulation thus become porous, inducing an epistemological crisis that requires not merely a regulatory update, but a conceptual reorientation.

3. The Race Between Regulation and Technological Innovation

The second tension, inextricably linked to the first, is of a practical-temporal and political nature. There is a structural misalignment, known as the *pacing problem*, between the linear and deliberative temporality of the legislative process and the exponential and disruptive pace of technological development. This discontinuity renders any attempt at specific regulation, focused on a particular application or technology, an exercise destined for rapid obsolescence. This pushes the search for new, more agile and principle-based regulatory architectures, but it also opens a field of geopolitical contention from which regulatory models emerge that reflect radically different governing philosophies. The European Union, with its AI Act, has adopted a human-centric and precautionary approach based on a hierarchical classification of risk (unacceptable, high, limited, minimal). This model aims to protect fundamental rights *ex ante* by imposing stringent obligations on high-risk systems and aspires to export its standards globally through the so-called "Brussels effect". The United States, in contrast, favours a decentralised, sector-specific, and market-oriented philosophy of "permissionless innovation". This encourages innovation without prior authorisation, entrusting risk management to self-regulation and *ex post* intervention by existing agencies (such as the FTC), in order not to stifle its technological leadership in a context of global competition. Finally, China presents a third, distinctly state-centred model, oriented toward social control and the achievement of national strategic objectives. Its governance is based on targeted regulations for specific sectors (such as recommendation algorithms or generative AI) and a mandatory registry of algorithms, pursuing a dual objective of technological primacy and internal stability. Legislating on AI today is therefore akin to trying to draw maps of a territory undergoing constant seismic shifts and expansion, a territory where the competition between geopolitical actors and large private corporations comes alive and intersects.

The ongoing transformation is further complicated by the fact that artificial intelligence, perhaps more than any previous technology, is not merely an object to which regulations are applied; it influences the normative processes themselves, reconfiguring how law and public opinion are produced. This occurs not only through the use of AI as a tool but due to the very nature of these particular technologies.

The instrumental use of AI to influence the public sphere is manifested in emblematic examples such as electoral manipulation through deepfakes or the algorithmic amplification of specific political messages. The latter dynamic, in particular, can emerge

as the consequence of an intentional action, but also as the unintended outcome of an intrinsic property of platforms optimised to maximise user engagement, ultimately leading to the systematic favouring of the communicative styles of certain movements. On top of this, a growing source of concern is the use of Large Language Models (LLMs) in the drafting of legal texts, which introduces a logic of automation into a domain ideally governed by human deliberation and interpretation. In parallel, the very nature of AI models, trained on vast data archives, means they inherit and amplify the biases and stereotypes present in society, risking the crystallisation of existing inequalities. Finally, the statistical and predictive nature of contemporary AI models promotes a transformation in the way politics itself is conceived, producing an "algorithmic governmentality" that privileges predictive and automated management over democratic deliberation.

4. AI as object and subject

In this scenario, AI ceases to appear as a mere tool and shows itself as an infrastructural technology that profoundly affects the informational and decision-making environment. Faced with such a complex set of challenges, the inclination toward technophobia or, conversely, the embrace of technological fatalism, though seemingly antithetical, both lead to similar dead ends, incapable as they are of confronting the current evolving landscape. It is, instead, increasingly necessary to inhabit this complexity by adopting a critical posture that replaces hasty generalisations with rigorous analysis rooted in concrete practices and harms. To this end, it is important to map the new forms of "harm" that are emerging, stratified into different layers and dimensions of the transformations at hand, which here, for illustrative reasons, we can divide into two orders: that of *social* harm and that of *epistemic* harm.

The first order of harm, the social one, manifests in how automated classification systems can reinforce and conceal, under a veneer of scientific objectivity, existing inequalities and pre-existing power relations in the contexts where these systems are deployed. Paradigmatic examples include the COMPAS algorithm in the U.S. judicial system, which exhibited racially disparate false positive rates; the childcare benefits scandal in the Netherlands, where dual nationality unexpectedly became a key indicator of fraud and disproportionately harmed families of immigrant origin[6]; or Optum's healthcare algorithm, which underestimated the health needs of Black patients by using past healthcare costs as a proxy for future need.

The second, more subtle order of harm is epistemic, which extends to forms of epistemic injustice. Generative AI acts as a powerful multiplier of such injustices: *testimonial* injustice, when models trained on corpora steeped in stereotypes undermine the credibility of entire social groups, and *hermeneutical* injustice, when the homogenising logic of the models erases minority narratives and culture. This culminates in the corrosion of collective trust in the public sphere, evidenced by the steady decline of trust in the news[7], and in the phenomenon of the "liar's dividend", whereby the very existence of deepfakes (which are

[6] This refers to the "toeslagenaffaire", the Dutch scandal regarding the algorithmic management of childcare benefits, which unjustly accused thousands of families of fraud.

[7] See N. Newman et al., *Reuters Institute Digital News Report 2025*, Reuters Institute for the Study of Journalism, Oxford 2025, p.73. The report highlights that overall trust in news has fallen further, settling at 26% globally in 2025 from 38% in 2018.

increasingly easy to create not only as static images but also as videos) allows public figures to deny the veracity of authentic evidence, undermining the foundations of political accountability. Here, the threat is not just falsehood, but the impossibility of establishing a shared, recognised access to reality that is fundamental to the existence of the public sphere as a shared space of visibility and discussion.

It is therefore necessary to develop critical thinking capacities that do not merely chase after technology in a perennial reactive struggle, but that anticipate its trajectories and possess the audacity to forge new conceptual categories. Addressing the tensions posed by AI today requires modes of thought that reaffirm the primacy of ethical judgment and political deliberation in the face of the logic of optimisation and profit at all costs. The challenge that awaits us is not merely technical or legal in nature; it is, in its most profound and urgent sense, philosophical, ethical, and political.

5. Issue Overview

The first section, "Inquiries", presents a series of reflections and investigations over broad, scenario-defining questions. Taken together, these essays draft a tentative geography of the "unstable grounds" upon which AI regulation is to be established. Alexei Grinbaum, in this Italian translation of a recent original essay, draws from philosophical and theological sources to articulate a suggestive look into the unprecedented questioning of temporality and responsibility inaugurated by the recent new wave of progress in generative AI technologies. Michelle Worthington explores the crucial nexus between the corporate and socio-economic structures that drive the growth and release of AI systems and the normative problem of responsibility: it is only by looking at the irresponsible dynamics that internally inhabit these emerging corporate subjects that we can correctly locate the genuine core of the regulation endeavour. The questioned centrality of the human is another fundamental area of the regulatory problem that is analysed in the article presented by Federico Reggio. In this perspective, human vulnerability is central, and the structural ambivalence of technological developments needs to be urgently acknowledged: this purpose, Reggio argues that restorative ethics, born in a different context, has much to teach about how we should approach the normative task when confronted with the risk of great harm that comes with the current shifts and changes. Lydia Farina and Anna-Maria Piskopani lead us into the algorithmic inner workings of AI systems and their problematic implications in terms of justice; the authors suggest that while much discourse is currently focusing on bias being hidden and replicated by AI, there is also a case to be made for algorithmic processing to be used as a powerful resource to diagnose and monitor bias. The relationship between justice and algorithms is also central in the final article of this section, by Enea Lombardi. Here the author seeks to integrate current approaches to algorithmic fairness with a more structural understanding of injustice: drawing inspiration from Iris M. Young's sensitivity towards the interplay of behaviours, norms, and institutions, Lombardi brings again our attention to the deep connection between AI development and the broader socio-historical context within which the ambivalent potential of algorithmic technologies is being deployed.

The second section, "Researches", offers a selection of articles that put their focus on narrower topics in AI regulation, by looking into specific fields of practices or by adopting a specific conceptual lens. These closer looks into the topic are not less relevant and allow

for a ground-level exploration of some emerging problems without forgetting the wider philosophical questions. Sung-Yeop JO offers an examination of the attribution of consciousness to AIs through an original application of Immanuel Kant's conceptual repertoire: the preliminary conclusion is that, while to some limited extent AI systems may be considered 'conscious' in a Kantian sense, they do not qualify within the same framework as a person in an ethical and legal sense. In the next paper, Siobhain Lash suggests an unexpected and stimulating intersectional analysis of the regulatory debate on abortion and the new emerging debates on Autonomous Vehicle Regulation in the United States: apparently distant, these two areas of normative controversy are currently traversed by similar concerning tendencies when it comes to the restrictions on bodily autonomy and freedom of mobility, paired with increasing levels of surveillance for women and marginalized and minority communities. Pier Francesco Miccichè looks into the role of librarians as mediators of our access to knowledge and advocates that the substitution of human intermediators with AI systems can be problematic because of their hidden biases and opaque outputs; the use of AI in this field should be rather directed, he argues, towards a fruitful integration into knowledge ecosystems that does not sideline librarians, but rather empowers them as educators and critical mediators. The contribution written by Alberto Romele, Dario Rodighiero, and Sabina Rosenbergova focuses on the visual representation of AI in stock image repositories: the analysis points out that the overabundant visual lexicon of particles, pixels, or voxels suggests an increasing datafication of the worldview of which AI is but one example in a more general trend that involves architecture and design. Finally, Patrizia Natale articulates a critical analysis of the introduction of AI in schools, highlighting a wide array of risks that range from the introduction of subtle and pervasive biases to the violation of privacy - especially of minors - to the spread of security threats. AI will likely have an increasing and impactful presence in the landscape of education, but two moves are recommended to minimise its negative implications: ethical AI design through participatory approaches that actively involve schools and educators, and a push towards digital education and critical awareness of technology in school curricula.