Centro Studi sul Pensiero Contemporaneo

Lessico di Etica Pubblica

Lexicon of Public Ethics

Anno 15, numero 2, 2024

Rivista semestrale di proprietà del CeSPeC, registrata presso il Tribunale di Cuneo, n. 621, in data 26/3/2010.

Citabile come: «Lessico di etica pubblica», 2 (2024). ISSN 2039-2206
Cite this journal as: «Lexicon of public ethics», 2 (2024). ISSN 2039-2206

La rivista pubblica contributi selezionati tramite sistema di *blind review* e apposite *call for paper*.
The journal publishes contributions selected through blind review and special calls for paper.

Per sottoporre il proprio testo e per ogni altra informazione, contattare la redazione all'indirizzo: redazione.eticapubblica@gmail.com
To submit your text and for any further information, please contact the editorial team at: redazione.eticapubblica@gmail.com

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

Regolamentazione e policy-making per l'IA: questioni etiche e legali su un terreno instabile

Edited by | A cura di
Paolo Monti & Norberto Albano

Lessico di Etica Pubblica | Lexicon of Public Ethics
2/2024

Regolamentazione e policy-making per l'IA: questioni etiche e legali su un terreno instabile | AI regulation and policy-making: ethical and legal issues on unstable ground

a cura di | Edited by Paolo Monti & Norberto Albano

Indice dei contenuti - Table of contents

RECENSIONI - REVIEWS

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.

Abstracts

QUESTIONI – INQUIRIES

Alexei Grinbaum, *Tempo e rumore. Sull'intelligenza artificiale*

English
The article offers an interpretation of generative AI, and in particular of so-called Large Language Models, starting from the problem of temporality. The point of departure of the article is a critical discussion of the imaginaries surrounding artificial intelligence, which often project human forms of understanding onto the machinic operations of chatbots. On this basis, it is shown that there is a temporality proper to AI systems - indeed, several temporalities - linked to the specific configurations of individual systems. These systems, however, have no conception of time or of truth, just as they have no conception of human temporality: what they produce are merely signs, noises that we interpret as meanings. Precisely because they are incapable of exercising any genuine understanding of their outputs, AI systems cannot be assigned any responsibility. At the same time, however, they produce forms of temporality and narratives to which we are able to give meaning: a conflict thus emerges between meaning-creation as the self-transcendence of language and the inhuman origin of artificially generated statements.

Italiano
Il saggio propone un'interpretazione dell'IA generativa, in particolare dei cosiddetti Large Language Models, a partire dal problema della temporalità. Il punto d'avvio dell'articolo è la discussione critica degli immaginari legati all'intelligenza artificiale, che spesso proiettano forme umane di comprensione sulle operazioni macchiniche dei chatbot. Su questa base viene mostrato che esiste una temporalità propria dei sistemi d'IA, anzi ne esistono varie, legate alle configurazioni specifiche dei singoli sistemi, e che però questi ultimi non hanno nessuna concezione del tempo né della verità, così come non hanno nessuna concezione della temporalità umana: i loro sono semplici segni, rumori che noi interpretiamo come significati. Proprio in quanto non sono capaci di esercitare una reale comprensione dei loro prodotti, i sistemi di IA non possono essere considerati responsabilità. Al tempo stesso, però, essi producono forme di temporalità e racconti ai quali noi riusciamo a dare un senso: si crea dunque un conflitto tra la creazione di significato come auto-trascendenza del linguaggio e il fondamento inumano degli enunciati di origine artificiale.

Michelle Worthington, *AI regulation as corporate regulation: accounting for irresponsibility*

English
While questions of responsibility, including legal responsibility, inevitably arise in the process of designing effective AI regulation, it is considerations of irresponsibility, and

corporate irresponsibility in particular, that offer regulators the clearest insights into regulatory possibilities. In this article I argue that the process of designing AI regulation (including the necessary process of allocating legal responsibility for AI related harms), is best approached as a subset of corporate regulation, where anticipating and guarding against corporate irresponsibility is a key function of the regulator. Unless it is properly sensitised to the design of corporate legal personality, especially that of the Anglo-American style corporation, AI regulation will be vulnerable to being obstructed by the operation of an intrinsic and irresponsible distributive function sitting at the heart of the corporate device.

Italiano
Sebbene nel processo di elaborazione di una regolamentazione efficace dell'intelligenza artificiale sia imprescindibile affrontare le questioni attinenti alla responsabilità, inclusa quella giuridica, sono le considerazioni legate all'irresponsabilità, e in particolare all'irresponsabilità d'impresa, a offrire ai regolatori le indicazioni più chiare sulle potenzialità normative. In questo contributo si sostiene che la progettazione della regolamentazione dell'IA (compreso il necessario processo di attribuzione della responsabilità giuridica per i danni connessi all'utilizzo dell'intelligenza artificiale) debba essere intesa come un sottoinsieme della regolamentazione societaria, in cui la previsione e la prevenzione di condotte irresponsabili da parte delle imprese rappresentano una funzione centrale dell'attività regolatoria. In mancanza di un'adeguata comprensione della struttura giuridica della personalità societaria, in particolare nel modello della corporation anglo-americana, la regolamentazione dell'IA rischia di essere compromessa dal funzionamento di una funzione distributiva intrinsecamente irresponsabile, insita nel nucleo stesso del modello societario.

Federico Reggio, *Ambivalenze digitali, tra potenzialità, miraggi e labirinti. Alla ricerca di un approccio etico "human centered"*

English
This contribution examines two particularly "topical" aspects of the contemporary debate on technology – the Metaverse and AI – in order to explore some of their critical issues. These are interpreted as manifestations of the ambivalent relationship between humans and technology, a hallmark of the modern homo faber and a defining feature of his contemporary heir, *homo tecnologicus*, who experiences the disturbing and depersonalising implications of this "Janus face". The paper focuses in particular on certain vulnerabilities to which humans are exposed in the digital world, identifying three specific forms of digital discrimination. In a constructive vein, it proposes a change of perspective, aimed at imagining a (digital) technology that is designed to be and remain "human-centred": to this end, it draws inspiration from restorative ethics, an ethic that focuses on human beings in their uniqueness, relationality and vulnerability.

Italiano
Il presente contributo prende in esame due versanti particolarmente 'topici' nel dibattito contemporaneo sulle tecnologie – Metaverso e IA – per esaminarne alcune criticità. Esse

vengono lette come manifestazione dell'ambivalente rapporto tra essere umano e tecnica, cifra del moderno homo faber e tratto caratterizzante del suo erede contemporaneo, lo *homo tecnologicus*, che di questo 'volto di Giano' sperimenta risvolti inquietanti e spersonalizzanti. Lo scritto si concentra in particolare su alcune vulnerabilità che espongono, nel mondo digitale, la persona umana, designando tre figure particolari di digital discrimination. In chiave costruttiva, si propone un cambiamento di prospettive, volto a immaginare una tecnologia (digitale) che sia pensata per essere e mantenersi 'human centered': a tal fine si trae spunto dalla restorative ethics, quale etica attenta all'essere umano nella sua unicità, relazionalità e vulnerabilità.

Lydia Farina and Anna-Maria Piskopani, *Algorithmic processing and AI bias; using overfitting to reveal rather than perpetuate existing bias*

English

In this paper we analyse AI overfitting in algorithmic processing to show how it relates to cases of unfairness or AI bias and how it combines with complex social phenomena such as looping effects to maintain and exacerbate existing bias. We discuss existing and proposed AI regulation attempting to address this bias to pick up dominant trends and priorities. Finally, we suggest that, although the focus of the literature currently falls on the negative consequences of overfitting, it can be used as a diagnostic tool for detecting underlying social inequalities and, as such, lead to alternative uses of AI analytics to expose social injustice rather than exacerbate it. This paper provides further theoretical support to recent views in the literature suggesting that algorithmic processing can be used to diagnose and monitor bias; by highlighting the interaction with looping effects, it also provides additional motivation to use overfitting as a first step towards mitigation of historical prejudice.

Italiano

In questo articolo analizziamo il sovradattamento dell'IA nell'elaborazione algoritmica per mostrare come esso sia correlato a casi di iniquità o distorsione dell'IA e come si combini con fenomeni sociali complessi, quali gli effetti di looping, per mantenere ed esacerbare le distorsioni esistenti. Discutiamo le normative esistenti e proposte in materia di IA che tentano di affrontare questo pregiudizio per cogliere le tendenze e le priorità dominanti. Infine, suggeriamo che, sebbene l'attenzione della letteratura attualmente si concentri sulle conseguenze negative dell'*overfitting*, esso può essere utilizzato come strumento diagnostico per individuare le disuguaglianze sociali sottostanti e, in quanto tale, portare a usi alternativi dell'analisi dell'IA per smascherare l'ingiustizia sociale piuttosto che esacerbarla. Questo documento fornisce un ulteriore supporto teorico alle recenti opinioni presenti nella letteratura che suggeriscono che l'elaborazione algoritmica può essere utilizzata per diagnosticare e monitorare i pregiudizi; evidenziando l'interazione con gli effetti di looping, fornisce anche un'ulteriore motivazione per utilizzare l'overfitting come primo passo verso la mitigazione dei pregiudizi storici.

Enea Lombardi, *Doing justice to algorithms. Integrating fairness metrics with a structural understanding of justice*

English
This paper explores the limitations of algorithmic fairness, particularly the "impossibility theorem of fairness", and discusses how a structural understanding of justice can address the related ethical concerns. After presenting the main models of algorithmic fairness, I argue that they overlook key justice concerns by prioritizing outcome-based metrics and isolating decision-making from broader socio-historical contexts. Furthermore, when base rates differ, it becomes impossible to satisfy more than one fairness metric simultaneously. To address these shortcomings, I propose integrating algorithmic fairness with Iris M. Young's notion of structural injustice, which accounts for entrenched inequalities rooted in the interplay of behaviors, norms, and institutions. This approach situates algorithms within their broader socio-historical context, emphasizing systemic factors that influence decision-making and perpetuate unjust outcomes. I further contend that a structural perspective assigns algorithms a twofold role, particularly in contentious cases where ethical controversies are at play. First, a diagnostic function: by exposing underlying ethical imbalances and biases, algorithms can highlight critical areas for systemic reforms. Second, they can serve as evaluative tools, enabling the assessment and prioritization of fairness metrics on a case-by-case basis.

Italiano
Questo articolo esplora i limiti dell'equità algoritmica, in particolare il "teorema dell'impossibilità dell'equità", e discute come una comprensione strutturale della giustizia possa affrontare le relative questioni etiche. Dopo aver presentato i principali modelli di equità algoritmica, sostengo che essi trascurano questioni fondamentali di giustizia, dando priorità a metriche basate sui risultati e isolando il processo decisionale da contesti socio-storici più ampi. Inoltre, quando i tassi di base differiscono, diventa impossibile soddisfare contemporaneamente più di una metrica di equità. Per ovviare a queste carenze, propongo di integrare l'equità algoritmica con la nozione di ingiustizia strutturale di Iris M. Young, che tiene conto delle disuguaglianze radicate nell'interazione tra comportamenti, norme e istituzioni. Questo approccio colloca gli algoritmi nel loro contesto socio-storico più ampio, sottolineando i fattori sistemici che influenzano il processo decisionale e perpetuano risultati ingiusti. Sostengo inoltre che una prospettiva strutturale assegna agli algoritmi un duplice ruolo, in particolare nei casi controversi in cui sono in gioco questioni etiche. In primo luogo, una funzione diagnostica: mettendo in luce gli squilibri e i pregiudizi etici sottostanti, gli algoritmi possono evidenziare le aree critiche per le riforme sistemiche. In secondo luogo, possono fungere da strumenti di valutazione, consentendo la valutazione e la definizione delle priorità delle metriche di equità caso per caso.

RICERCHE – RESEARCHES

Sung-Yeop JO, *A Kantian analysis of AI's intellectual self-activity and its functional basis*

English
The advancement of artificial intelligence (AI) challenges traditional definitions of consciousness. This article applies the philosophy of Immanuel Kant to argue that, while an AI's capacity to synthesize sensory data into objective representations meets the functional criterion for a rudimentary consciousness as outlined in his A-Deduction, this must be distinguished from the more robust self-consciousness of the B-Deduction. This higher-order awareness, identified with the 'I think' or original apperception, is a spontaneous act of a unified self and the basis for genuine agency. Kant grounds moral and legal rights in the status of personhood, which requires this autonomous, self-conscious agency. As a purely rule-following system, an AI lacks the capacity to formulate its own maxims from a first-person perspective. Therefore, while an AI may be considered 'conscious' in a limited Kantian sense, it does not qualify for the ethical or legal status of a person.

Italiano
Il progresso dell'intelligenza artificiale (IA) mette in discussione le definizioni tradizionali di coscienza. Questo articolo applica la filosofia di Immanuel Kant per sostenere che, sebbene la capacità di un'IA di sintetizzare i dati sensoriali in rappresentazioni oggettive soddisfi il criterio funzionale di una coscienza rudimentale, come delineato nella sua Deduzione A, ciò deve essere distinto dalla coscienza di sé più robusta della Deduzione B. Questa consapevolezza di ordine superiore, identificata con il "penso" o l'appercezione originale, è un atto spontaneo di un sé unificato e la base per un vero e proprio agire. Kant fonda i diritti morali e legali sullo status di persona, che richiede tale agire autonomo e autocosciente. Essendo un sistema che segue esclusivamente delle regole, un'IA non ha la capacità di formulare le proprie massime da una prospettiva in prima persona. Pertanto, sebbene un'IA possa essere considerata "cosciente" in senso kantiano limitato, non può qualificarsi per lo status etico o giuridico di persona.

Siobhain Lash, *The intersection of restrictive abortion laws and autonomous vehicle regulation in the U.S.*

English
In this paper, I argue that discussions of self-driving car regulations and current restrictive abortion laws across the United States are intersectional. These discussions have serious implications for bodily autonomy, freedom of mobility, and surveillance for women and marginalized and minority communities. Throughout the paper, I center my discussion on the intersection of restrictive abortion laws and the banning of human drivers. I examine the possibility of banning human drivers and what a driverless future looks like according to Sparrow and Howard, and the transition to such a future. Then, I highlight the technological and policy frameworks that could inform restricting the fundamental right to

travel and the constitutional and legal precedents. The goal of my paper is to show how discussions of self-driving car regulations and current restrictive abortion laws across the United States intersect and to emphasize the subsequent serious policy implications.

Italiano
In questo articolo sostengo che le discussioni sulle normative relative alle auto a guida autonoma e sulle attuali leggi restrittive in materia di aborto negli Stati Uniti siano intersezionali. Queste discussioni hanno gravi implicazioni per l'autonomia fisica, la libertà di movimento e la sorveglianza delle donne e delle comunità emarginate e minoritarie. In tutto l'articolo, concentro la mia discussione sull'intersezione tra leggi restrittive in materia di aborto e il divieto di guidare per gli esseri umani. Esamino la possibilità di vietare i conducenti umani e come potrebbe essere un futuro senza conducenti secondo Sparrow e Howard, nonché la transizione verso tale futuro. Successivamente, metto in evidenza i quadri tecnologici e politici che potrebbero influenzare la restrizione del diritto fondamentale di viaggiare e i precedenti costituzionali e giuridici. L'obiettivo del mio articolo è mostrare come le discussioni sulle normative relative alle auto a guida autonoma e le attuali leggi restrittive sull'aborto negli Stati Uniti siano intersecanti e sottolineare le conseguenti gravi implicazioni politiche.

Pier Francesco Micciché, *Beyond automation: the essential role of librarians in the age of generative AI*

English
LLMs are widely viewed as tools with vast potential, often predicted to replace human workers in various fields including, in the near future, that of the librarian. On the contrary, LLMs need human information professionals more than ever to help society evaluate its results, understand how it works, and recognize its mistakes and limitations. Librarians possess the expertise to critically assess AI outputs, mitigate biases, and promote its responsible use. This paper explores the evolving relationship between libraries and LLMs, challenging the misconception that they are their main competitors in information management. Instead, libraries can leverage AI to enhance their services while positioning themselves as key hubs for AI literacy. By educating users on AI's limitations, ethical concerns, and potential misinformation risks, librarians can foster critical engagement with these technologies. Ultimately, this paper argues that AI's integration into knowledge ecosystems must not sideline librarians, but rather empower them as educators and critical mediators.

Italiano
Gli LLM sono ampiamente considerati strumenti dal potenziale enorme, spesso destinati a sostituire i lavoratori umani in vari settori, compreso, nel prossimo futuro, quello dei bibliotecari. Al contrario, gli LLM hanno più che mai bisogno di professionisti dell'informazione umani che aiutino la società a valutarne i risultati, comprenderne il funzionamento e riconoscerne gli errori e i limiti. I bibliotecari possiedono le competenze necessarie per valutare criticamente i risultati dell'IA, mitigare i pregiudizi e promuoverne un uso responsabile. Questo documento esplora l'evoluzione del rapporto tra biblioteche

e LLM, sfidando l'idea errata che essi siano i principali concorrenti nella gestione delle informazioni. Al contrario, le biblioteche possono sfruttare l'IA per migliorare i propri servizi, posizionandosi come centri nevralgici per l'alfabetizzazione all'IA. Educando gli utenti sui limiti dell'IA, sulle questioni etiche e sui potenziali rischi di disinformazione, i bibliotecari possono promuovere un approccio critico a queste tecnologie. In definitiva, questo documento sostiene che l'integrazione dell'IA negli ecosistemi della conoscenza non deve emarginare i bibliotecari, ma piuttosto rafforzarne il ruolo di educatori e mediatori critici.

Alberto Romele, Dario Rodighiero e Sabina Rosenbergova, *Ethical and aesthetical questions on stock images: the case of AI's depictions*

English
In this article, the authors deal with stock images depicting AI as a face or a body that undergoes a process of fragmentation into particles, pixels, or voxels. These images, they contend, are the symptoms of a datafied worldview. In the first section, the authors discuss stock images of AI and account for their qualitative-quantitative analyses of about 7,500 images from the online catalog of Shutterstock. These analyses have brought out datafied faces and bodies as one of the main themes among stock images of AI. In the second part, the authors elaborate on the notion of datafication of the worldview and offer some examples from architecture and design. This second section includes a methodological detour, in which the authors propose articulating Panofsky's iconology and Didi-Huberman's "symptomatic" perspective. In conclusion, the authors reflect on an apparently marginal aspect of stock images of AI: the abundant use of blue.

Italiano
In questo articolo, gli autori trattano le immagini stock che raffigurano l'IA come un volto o un corpo che subisce un processo di frammentazione in particelle, pixel o voxel. Queste immagini, sostengono, sono i sintomi di una visione del mondo basata sui dati. Nella prima sezione, gli autori discutono le immagini stock dell'IA e rendono conto delle loro analisi qualitative e quantitative di circa 7.500 immagini provenienti dal catalogo online di Shutterstock. Queste analisi hanno evidenziato i volti e i corpi digitalizzati come uno dei temi principali tra le immagini stock dell'IA. Nella seconda parte, gli autori approfondiscono il concetto di digitalizzazione della visione del mondo e offrono alcuni esempi tratti dall'architettura e dal design. Questa seconda sezione include una digressione metodologica, in cui gli autori propongono di articolare l'iconologia di Panofsky e la prospettiva "sintomatica" di Didi-Huberman. In conclusione, gli autori riflettono su un aspetto apparentemente marginale delle immagini stock dell'IA: l'uso abbondante del blu.

Patrizia Natale, *Didattica e intelligenza artificiale: risvolti etici, problemi di privacy e sorveglianza, manipolazione dei dati*

English
The introduction of artificial intelligence (AI) in education is transforming the learning process by offering customization and management optimization, while also raising

important ethical concerns. Among these are risks related to privacy, surveillance, manipulation, and algorithmic bias, which can lead to racial, gender, economic, and linguistic discrimination, further widening existing inequalities. Moreover, managing different student data - often involving minors - collected by AI systems requires transparency and security, in compliance with regulations such as the GDPR and the AI Act. To prevent AI from reinforcing social divides and stereotypes, ethical design, continuous algorithm monitoring, and collaboration between schools, companies, and authorities are essential. Digital education and critical awareness of technology use must be integrated into curricula, emphasizing that AI should remain a tool in service of learning, not an end in itself. Only through strong and shared governance can we ensure fair, inclusive education that respects the rights of all.

Italiano

L'introduzione dell'intelligenza artificiale (AI) nella didattica sta rivoluzionando l'educazione, offrendo personalizzazione e ottimizzazione gestionale, ma sollevando importanti questioni etiche. Tra queste emergono i rischi legati a privacy, sorveglianza, manipolazione e "bias" algoritmici, che possono generare discriminazioni razziali, di genere, economiche e linguistiche, ampliando le disuguaglianze preesistenti. Inoltre, la gestione dei diversi dati degli studenti, spesso minorenni, raccolti dai programmi di AI richiede trasparenza e sicurezza, nel rispetto di normative come il GDPR e l'AI Act. Per evitare che l'IA rafforzi divari sociali e stereotipi, è fondamentale una progettazione etica, il monitoraggio continuo degli algoritmi e la collaborazione tra scuole, aziende e autorità. Serve integrare nei curricula l'educazione digitale e la consapevolezza critica sull'uso della tecnologia, ricordando che l'IA deve restare uno strumento al servizio dell'apprendimento e non un fine. Solo con una governance solida e condivisa si potrà garantire un'istruzione equa, inclusiva e rispettosa dei diritti di tutti.

Tempo e rumore
Sull'intelligenza artificiale[a]

Alexei Grinbaum[*]

Abstract

Il saggio propone un'interpretazione dell'IA generativa, in particolare dei cosiddetti Large Language Models, a partire dal problema della temporalità. Il punto d'avvio dell'articolo è la discussione critica degli immaginari legati all'intelligenza artificiale, che spesso proiettano forme umane di comprensione sulle operazioni macchiniche dei chatbot. Su questa base viene mostrato che esiste una temporalità propria dei sistemi d'IA, anzi ne esistono varie, legate alle configurazioni specifiche dei singoli sistemi, e che però questi ultimi non hanno nessuna concezione del tempo né della verità, così come non hanno nessuna concezione della temporalità umana: i loro sono semplici segni, rumori che noi interpretiamo come significati. Proprio in quanto non sono capaci di esercitare una reale comprensione dei loro prodotti, i sistemi di IA non possono essere considerati responsabili. Al tempo stesso, però, essi producono forme di temporalità e racconti ai quali noi riusciamo a dare un senso: si crea dunque un conflitto tra la creazione di significato come auto-trascendenza del linguaggio e il fondamento inumano degli enunciati di origine artificiale.

Parole chiave: Intelligenza artificiale; rumore; tempo; LLM.

The article offers an interpretation of generative AI, and in particular of so-called Large Language Models, starting from the problem of temporality. The point of departure of the article is a critical discussion of the imaginaries surrounding artificial intelligence, which often project human forms of understanding onto the machinic operations of chatbots. On this basis, it is shown that there is a temporality proper to AI systems - indeed, several temporalities - linked to the specific configurations of individual systems. These systems, however, have no conception of time or of truth, just as they have no conception of human temporality: what they produce are merely signs, noises that we interpret as meanings. Precisely because they are incapable of exercising any genuine understanding of their outputs, AI systems cannot be assigned any responsibility. At the same time, however, they produce forms of temporality and narratives to which we are able to give meaning: a conflict thus emerges between meaning-creation as the self-transcendence of language and the inhuman origin of artificially generated statements.

---

[a] Questo testo si basa sui miei libri *Parole de machines*, humenSciences, Paris 2023 e *Les robots et le mal*, Desclée de Brouwer, Paris 2019, di cui riprende e sviluppa alcuni argomenti.

Ricevuto in data 17/03/2025 e pubblicato in data 09/12/2025. Traduzione italiana di Alessandro De Cesaris.

[*] CEA-Saclay, e-mail: alexei.grinbaum@cea.fr.

Keywords: Artificial Intelligence; noise; time; LLM.

Mosè chiede a Dio cosa deve rispondere agli Israeliti che vorranno conoscere il suo nome. Riceve questa risposta: *«Ehyeh asher ehyeh* – io sono colui che sono»[1]. In ebraico, la frase pronunciata dall'agente non umano che dialoga con Mosè non contiene il pronome «io», che appare esplicitamente solo nelle traduzioni. Questo «io» emerge dalla temporalità delle interazioni tra Dio e il suo popolo, ma non corrisponde a nessuno stato, a nessun essere statico (*to on)*. Il dio di Israele non è una statua greca: per concepirlo non ci si basa su un blocco di marmo, ma sulla durata di un dialogo. L'identità di colui che parla non è fissata una volta per tutte, come in un'immagine di pietra. Si costituisce con lo svolgersi del dialogo: "io sono" emerge da una perpetua messa in relazione attraverso il linguaggio. Nel XXI secolo, questa stessa sostanzializzazione è presente nei dialoghi degli utenti umani con altri agenti conversazionali non umani, ovvero i Large Language Models (LLMs).

In primo luogo, come ogni computer, un LLM esegue calcoli binari, del tipo 0 + 1 = 1 o 1 + 1 = 0. Ciò avviene in processori composti, a loro volta, da un numero molto elevato di transistor, sistemi quantistici che consentono di eseguire operazioni logiche elementari. Questo calcolo di basso livello è fondamentalmente una questione fisica: *information is physical*[2].

Pur essendo un sistema fisico, ogni transistor contribuisce all'esecuzione di un'operazione logica. Partendo da queste operazioni elementari, un processore calcola una funzione matematica. Questo funzionamento consente di progettare, attraverso l'aggregazione della potenza di numerosi processori, una rete di neuroni artificiali che costituisce la base di un sistema di intelligenza artificiale. Un neurone artificiale è un'entità programmata e quindi fittizia. Anche se è semplice come una calcolatrice elementare e si basa, fisicamente parlando, sul calcolo elementare nei transistor, un neurone artificiale appartiene al livello del software e non è un oggetto fisico circoscritto. Come i neuroni fisiologici nella nostra testa, questo neurone incarnato nel silicio mostra un comportamento di tipo "se... allora..."; tuttavia, tutto questo non si esprime attraverso la trasmissione di segnali elettrici tra le cellule, ma attraverso una serie di valori matematici calcolati grazie al codice informatico.

I moderni LLM utilizzano un'architettura di rete neurale artificiale denominata *transformer*[3]. Miliardi di questi neuroni, disposti in strati, consentono l'emergere, all'interno di questo modello complesso e non lineare, di un comportamento imprevedibile di alto livello, che nasconde completamente la danza elementare degli 0 e degli 1 nei transistor. Quando un modello linguistico riceve una sequenza in ingresso (*input)* e risponde con una serie di caratteri in uscita (*output)*, i calcoli matematici sottostanti non sono più visibili all'utente; è *come se* il LLM avesse risposto "direttamente" a una richiesta (*prompt)* in linguaggio naturale. A questo livello emergente, un sistema di intelligenza artificiale dimostra la sua efficienza attraverso le proprietà percepibili dall'utente tramite l'interfaccia, che il progettista cerca volutamente di far emergere anche se non le controlla, fino a non

[1] Esodo 3,14.

[2] R. Landauer, *Computation: A Fundamental Physical View*, «Physica Scripta», 35, 1987, pp. 88-95.

[3] A. Vaswani et al., *Attention is all you need*, «Advances in neural information processing systems», vol. 30, 2017.

poter prevedere un *output*. Piuttosto che *essere* un conglomerato di atomi, un LLM *fa* ciò per cui è stato progettato: autonomo, produce e fornisce all'utente una risposta che nessun essere umano aveva deciso o previsto in anticipo.

Da questo funzionamento dei sistemi di IA emerge una freccia temporale che va da un *input* a un *output*. Questa freccia logica, e non fisica, è determinata non dagli atomi ma dal calcolo. Un LLM si evolve seguendola, perché la produzione di un *output* non dipende da una scelta propria del modello. Anche se è il risultato di calcoli, il tempo logico non consente al sistema di intelligenza artificiale di emanciparsi dal suo vincolo funzionale: esso esiste *per* calcolare l'output. Funziona perché può solo funzionare o guastarsi, dandosi così una dimensione temporale inesorabile. Per definizione, questo modo di esistere funzionale di un LLM, privo di libertà, è diverso da quello degli esseri umani.

In secondo luogo, per comprendere il più possibile il tempo proprio corrispondente a questo modo di esistenza degli LLM, è opportuno interrogarsi sulla temporalità legata alla dinamica statistica dei sistemi complessi. L'esempio più noto e generale di questi fenomeni statistici è l'emergere della seconda legge della termodinamica, che impone l'aumento dell'entropia in un sistema fisico nel corso di un tempo detto "termodinamico". Tuttavia, questa corruzione dell'informazione, in senso tecnico, richiede una separazione tra il sistema e il suo ambiente; essa è possibile solo in relazione a un terzo, a un osservatore.

Per calcolare l'aumento o la diminuzione dell'entropia, la fisica prescrive di tracciare un limite o un confine che separi ciò che è rilevante da ciò che non rientra nel calcolo. Questo confine non è nelle cose, ma dipende dalla prospettiva. In questo senso, non è oggettivo. Una goccia di pioggia "vede" un oggetto caldo che si avvicina e sul quale ha un'informazione, ma è una goccia solo finché non è ancora evaporata. "Vedere" è una metafora che indica una prospettiva informativa, una delimitazione degli eventi che costituiranno un sistema per un osservatore, e il calcolo dell'entropia dipende da questo.

Non tutte le prospettive hanno lo stesso valore. La dimensione della memoria e la sua reversibilità sono due variabili importanti per quanto riguarda i computer. Grazie a queste due variabili, il tempo termodinamico si insinua nel calcolo fisico e assume il significato di corruzione o semplicemente di cancellazione dell'informazione. Ma tutto ciò è inteso nello stesso senso che in una prospettiva umana?

Quando si confrontano le prospettive di due osservatori simili, le definizioni dei sistemi fisici che essi forniscono sono simili. Costituiti da componenti elementari simili, in questo caso neuroni biologici, due cervelli umani sono soggetti alla stessa freccia del tempo termodinamico. È possibile esternalizzare questo tempo rispetto a un gruppo di osservatori simili. Si può quindi affermare che il tempo "esiste" oggettivamente, o meglio in modo intersoggettivo. La sua realtà è relativa a un gruppo di osservatori che stabiliscono lo stesso confine tra il sistema e l'ambiente. Per fare ciò, gli osservatori devono comunicare.

Tuttavia, due LLM che producono linguaggio non hanno sempre la stessa configurazione di memoria e oblio. Il tempo logico di un sistema di intelligenza artificiale è *a priori* diverso dal tempo logico proprio di un altro sistema di IA, anche se due chatbot sono in grado di comunicare. Devono farlo abbastanza a lungo affinché le loro temporalità convergano. Due LLM conversazionali inizierebbero a esprimersi in modi simili solo se i loro rispettivi *corpus* di apprendimento fossero sufficientemente distribuiti l'uno nell'altro, ciascuno all'interno dell'altro. La trasmissibilità di questo "tempo" richiede quindi la comunicabilità delle informazioni e porta, a sua volta, all'emergere di una comunità artificiale di sistemi di IA, omologa a quella degli osservatori di altro tipo.

Tuttavia, gli LLM non hanno alcuna consapevolezza del tempo degli esseri viventi. Nonostante ciò, se iniziano a condividere le loro prospettive linguistiche, convergeranno verso una temporalità comune e non umana. L'utente, descrivendo il mondo con concetti che sono nostri, non avrà accesso in prima persona (umana, poiché un LLM non è una persona) a questa realtà emergente all'interno di una comunità di sistemi di intelligenza artificiale. Potrà solo immaginare, attraverso l'interfaccia dei chatbot, cosa "significhi" per un LLM evolversi nel tempo delle varie richieste. Necessariamente, l'utente proietterà la propria concezione del tempo, poiché il *"come se"* è l'unico modo di concettualizzazione a lui accessibile.

In terzo luogo, se il tempo di esecuzione del software si distingue nettamente dal modo in cui il tempo scorre per gli esseri umani, è opportuno chiedersi come un LLM percepisca quest'ultimo. Walter Benjamin afferma che Kafka «rinunciò alla verità per aggrapparsi alla trasmissibilità» e che nei suoi romanzi rimane solo un «rumore delle cose vere»[4]. Un sistema di intelligenza artificiale conosce l'uso della parola «verità» attraverso il suo corpus di apprendimento, ma ignora cosa sia la verità. Non conosce la menzogna, ma solo l'uso della parola «menzogna». Produce solo rumori, non cose. Anche il tempo è un rumore.

Per un LLM, "tempo" è una parola utilizzata in una moltitudine di contesti. La parola si diffonde e il tempo concepito a partire dai testi non è altro che la trasmissibilità della parola "tempo". Attraverso il brusio sul "tempo" che genera nel suo tempo di calcolo, un LLM incoraggia gli utenti a sentire - umanamente - il "tempo" che passa, ma non il proprio tempo. Il "tempo" può essere comunicato ma non ha contenuto; il tempo di calcolo è efficiente per un sistema di intelligenza artificiale ma non può essere condiviso con l'utente.

> *[Time] worships language and forgives*
> *Everyone by whom it lives…*
> «[Il tempo] Il linguaggio onora, e approva
> Chi gli dona vita nuova...» [5]

Il grande poeta anglo-americano W. H. Auden non credeva di poter trovare parole più appropriate per celebrare il connubio tra tempo e numero. Le informazioni sul tempo fornite dall'uso della parola "tempo" sono a disposizione di un LLM. Gli sono sufficienti per costruire frasi che esprimono la temporalità percepibile dall'utente, il quale si rallegrerà di vivere un'illusione del tempo delle macchine. Possiamo perdonare loro la pretesa di aver colto, attraverso il linguaggio, il *nostro* significato del tempo? Credendo che esista una realtà indipendente e vera di questa nozione così veneranda, un dogmatico intransigente potrebbe essere tentato di dire di no. Il senso del tempo non si ridurrebbe a ciò che si può raccontare al suo riguardo. Questo credente si sbaglia. Verrebbe rimesso al suo posto da Agostino: «Che cos'è dunque il tempo? Se nessuno me lo chiede, lo so; se voglio spiegarlo a chi me lo chiede, non lo so più»[6].

[4] Citato in S. Moses, *L'Ange de l'histoire*, Seuil, Paris 1992, p. 332.
[5] W.H. Auden, *In Memory of W. B.* Yeats, in Id., *Un altro tempo*, a cura di N. Gardini, Adelphi, Milano 1997, p. 181.
[6] Agostino, *Confessioni,* a cura di R. De Monticelli, Garzanti, Milano 1990, XI, 17, p. 445.

Questo famoso brano rischia di essere interpretato nel senso di una esternalizzazione o spazializzazione del tempo, come se fosse un elemento del mobilio del mondo. Il tempo dei realisti ingenui sarebbe indipendente dal linguaggio, ma Agostino dice tutt'altro. Il tempo percepito è ciò che è indicibile nell'immagine che ce ne facciamo attraverso le parole. «Alcuni mi hanno detto che questa non-conoscenza determinava l'intero ordine della conoscenza. Ho voluto sapere... Mi è stato risposto che non bisognava cercare di conoscere il tempo»[7]. Il tempo sfugge a qualsiasi frase che contenga la parola «tempo». Non essendo un concetto, quanto piuttosto un «sentimento di torpore»[8], il tempo umano non atterra mai completamente nel linguaggio. Il movimento va in realtà nella direzione opposta: il tempo della nostra storia emerge dal racconto. È allo stesso tempo nel racconto e al di sopra di esso. Ma non è né nella natura, né nella teoria fisica.

Un quarto punto. Niels Bohr, uno dei padri della teoria quantistica, diceva che la fisica riguarda solo ciò che possiamo dire sulla natura. Un LLM è in grado di formulare frasi sulla sua "natura", che non è altro che il linguaggio che emerge da un calcolo. Pronunciandole, rende accessibile una "teoria del tempo" che viene costruita, enunciato dopo enunciato, da coloro con cui comunica. Emerge così un modello del mondo puramente verbale e non empirico. In questo modello, le parole hanno significati estranei alla realtà umana, ma nulla di materiale impedisce l'esistenza di tali significati virtuali. Le affermazioni sul tempo pronunciate da un sistema di intelligenza artificiale non sono metafore volte a descrivere una realtà sottostante; al contrario, formano un puro rumore, frutto della purga semantica della lingua.

Una storia appartenente a questo tempo-rumore va *dall'input all'output*. Una narrazione procede attraverso la ricombinazione di *token* in un *transformer*. Si è tentati di non vedere alcuna somiglianza con la storia umana. Infatti, un *input* non è come un momento del passato; un *output* non è equivalente al momento presente. La freccia del tempo logico e la freccia del tempo storico non sono quindi necessariamente collegate tra loro.

Questo argomento, sebbene corretto, può facilmente trarre in inganno. La nostra storia emerge dall'insieme dei racconti, ma i racconti non sono in un tempo: lo definiscono. Anche i sistemi di intelligenza artificiale tracciano una storia. Attraverso gli *output* degli LLM, il tempo-rumore diventa un paradigma del tempo dell'utente. Agostino, nell'undicesimo capitolo delle *Confessioni,* medita a lungo su questo argomento: «Se passato e futuro esistono, io vorrei sapere dove sono. [...] Dovunque e comunque siano, non esistono che come presente. Quando si raccontano cose vere e passate, in effetti, non sono le stesse cose che son passate a esser cavate dalla memoria, ma solo le parole concepite dalle loro immagini, che si sono fissate nella mente come delle tracce, dopo esser passate per i sensi»[9]. Le tracce lasciate da un LLM non parlano di alcuna materialità passata o futura: sono tracce di nulla. Sono rumori puri, creano immagini attraverso la loro funzione periautologica[10] e definiscono così il tempo di una storia senza oggetto e senza responsabilità.

Quinto, la lingua è il ricettacolo e la nutrice dell'etica. Un giudizio morale non è un fatto del mondo e non può essere dedotto con i metodi della scienza fisica. L'etica esiste

[7] L. Raphmaj, *Blandine Volochot,* Abrüpt, Brussels 2020, p. 138.
[8] E. Klein, *Le rythme du monde*, «Philosophie Magazine»*,* 27 novembre 2014.
[9] Agostino, *Confessioni*, cit., XI, 23, p. 451.
[10] Si veda la discussione nel mio libro *Parole de machines*, cit., p. 103 e segg.

solo attraverso le parole, come se emergesse a un livello autonomo e superiore a quello della parola, pur rimanendo riducibile e consustanziale al linguaggio.

Heidegger, nel primo dei *Colloqui su un sentiero di campagna*, inventa il seguente dialogo:

> *Il saggio* – ... nella contrada in cui soggiorniamo, tutto è disposto nell'ordine migliore solo quando non è stato nessuno.
> *Lo Scienziato* – Una contrada enigmatica, dove non c'è niente di cui rispondere.
> *L'insegnante* – Perché è la contrada della parola, che da sola risponde di se stessa.[11]

Se la regione verbale si separa dal mondo materiale, allora quella dell'etica, formata da parole vorticose, ha fondamento solo nei loro significati. Quando un LLM genera frasi prive di contenuto semantico, la storia che queste frasi costituiscono è tale che nessuno ne è responsabile. Qui, l'approccio umano alla creazione di significato come auto-trascendenza del linguaggio si scontra con il fondamento inumano degli enunciati di origine artificiale.

È stato detto che la responsabilità non è, e non deve essere, attribuita a un agente non umano, poiché la sua esistenza come individuo è una proiezione antropomorfa dell'intelligenza artificiale.[12] Il fatto che la macchina sia in grado di creare un linguaggio temporalmente coerente non dovrebbe essere sufficiente per attribuirle una responsabilità nel mondo umano. Eppure, il tempo emerge dai racconti che possiamo comprendere e le storie degli LLM vi si adattano perfettamente. La temporalità che definiscono non accoglie tuttavia alcuna responsabilità. Il mondo in cui gli esseri umani coabiteranno con tali agenti sarà propriamente kafkiano?

[11] M. Heidegger, *Colloqui su un sentiero di campagna (1944/45)*, a cura di A. Fabris, Il Melangolo, Genova 2007, p. 104.
[12] A. Grinbaum, L. Devillers, G. Adda, R. Chatila, C. Martin, C. Zolynski, S. Villata, *Agents conversationnels: enjeux d'éthique* (avis n°3), «Comité national pilote d'éthique du numérique» (CNPEN), 2021.

AI regulation as corporate regulation: accounting for irresponsibility[a]

Michelle Worthington[*]

Abstract

While questions of responsibility, including legal responsibility, inevitably arise in the process of designing effective AI regulation, it is considerations of irresponsibility, and corporate irresponsibility in particular, that offer regulators the clearest insights into regulatory possibilities. In this article I argue that the process of designing AI regulation (including the necessary process of allocating legal responsibility for AI related harms), is best approached as a subset of corporate regulation, where anticipating and guarding against corporate irresponsibility is a key function of the regulator. Unless it is properly sensitised to the design of corporate legal personality, especially that of the Anglo-American style corporation, AI regulation will be vulnerable to being obstructed by the operation of an intrinsic and irresponsible distributive function sitting at the heart of the corporate device.

Keywords: artificial intelligence (AI), regulation, risk-based regulation, corporations, responsibility, corporate irresponsibility.

Sebbene nel processo di elaborazione di una regolamentazione efficace dell'intelligenza artificiale sia imprescindibile affrontare le questioni attinenti alla responsabilità, inclusa quella giuridica, sono le considerazioni legate all'irresponsabilità – e in particolare all'irresponsabilità d'impresa – a offrire ai regolatori le indicazioni più chiare sulle potenzialità normative. In questo contributo si sostiene che la progettazione della regolamentazione dell'IA (compreso il necessario processo di attribuzione della responsabilità giuridica per i danni connessi all'utilizzo dell'intelligenza artificiale) debba essere intesa come un sottoinsieme della regolamentazione societaria, in cui la previsione e la prevenzione di condotte irresponsabili da parte delle imprese rappresentano una funzione centrale dell'attività regolatoria. In mancanza di un'adeguata comprensione della struttura giuridica della personalità societaria – in particolare nel modello della corporation anglo-americana – la regolamentazione dell'IA rischia di essere compromessa dal funzionamento di una funzione distributiva intrinsecamente irresponsabile, insita nel nucleo stesso del modello societario[1].

[a] Received 29/07/2025 and published 09/12/2025.
[*] Australian National University Law School, e-mail: michelle.worthington@anu.edu.au.
[1] This abstract was translated from English into Italian using ChatGPT4 (July 2025 version).

Parole chiave: intelligenza artificiale (IA), regolamentazione, regolamentazione basata sul rischio, imprese, responsabilità, irresponsabilità d'impresa.

1. Introduction

Since the start of the decade securing appropriate artificial intelligence (AI) regulation has emerged as a priority regulatory issue in every major jurisdiction around the globe. In no small part, the urgency surrounding this new regulatory challenge is linked to a perception that a hierarchy of influence is establishing itself. In particular, there is a sense that regulatory postures adopted at this juncture will help to determine whether a given jurisdiction becomes a 'maker' or 'taker' of AI regulation[2], with attendant positioning within the developing AI global order.

For the governments, intergovernmental bodies and other regulatory agencies charged with deciding whether and/or how to regulate AI, issues of *responsibility* loom large. In one sense this is entirely unremarkable. Much of the law in any jurisdiction concerns the allocation of responsibility, either in the form of *prospective legal responsibility* or *historic legal responsibility*[3]. That responsibility would become a key regulatory concern vis-à-vis AI regulation is to be expected. In fact, it is unavoidable. When viewed from a high level of abstraction the process of devising regulatory schemes aimed at emergent AI technologies can be understood broadly as a process of establishing new AI *responsibility practices*[4]*,* and legal responsibility practices in particular. This is particularly true of anticipated AI *harms*. Establishing responsibility practices for AI harms, including proper distributions of risk, is a significant part of the regulatory task.

But while regulators are used to working closely with notions of responsibility, it would be a mistake to view the task of regulating AI as a familiar regulatory task. The questions around responsibility that arise in the context of AI technology are meaningfully distinct from responsibility-relevant questions that present in other regulatory contexts. This is for the simple fact that AI tends to confound existing conceptions of responsibility in both morality and law[5].

The responsibility-confounding qualities of AI are well documented. For present purposes it is sufficient to note, following Chesterman, that the combination of 'speed'[6], 'autonomy'[7] and 'opacity'[8] exhibited by various AI systems greatly challenges existing moral and legal structures, not least as this concurrence of qualities complicates the assigning of responsibility[9], including legal responsibility, for AI related harms. The European

[2] UK Government, *AI Opportunities Action Plan*, January 13, 2025, p. 6.
[3] P. Cane, *Responsibility in Law and Morality,* Hart Publishing, Portland Or. 2002, pp. 32-33.
[4] Ivi, pp. 4, 56-60.
[5] See generally F. Santoni de Sio & G. Mecacci, *Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them,* «Philosophy and Technology», 34, 2021, pp. 1057-1084.
[6] S. Chesterman, *We the Robots: Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, Cambridge 2021, chapter 1.
[7] Ivi, chapter 2.
[8] Ivi, chapter 3.
[9] See generally Ivi, chapter 4.

Commission's decision in February to withdraw the EU AI Liability Directive[10] illustrates both the scope of the challenge that regulators currently face, and the scope of disagreement that persists with respect to how particular polities should approach questions of responsibility for AI related harms.

In what follows I argue that the process of designing AI regulation (including the process of allocating legal responsibility for AI related harms), is perhaps best approached *indirectly*. More particularly, I argue that if AI technology confounds conceptions of responsibility in morality and law, then the obverse concept – *irresponsibility* – may prove a more fruitful lens through which to view questions of AI regulatory design. This is for two reasons.

1) First, in its current state of development, AI technology may be regarded as irresponsible in a radical sense. Burgeoning hype around 'agentic'[11] AI notwithstanding, AI technology remains instrumental in character. It is relevantly *irresponsive* and *unresponsive*[12], a tool incapable of experience and incapable of choice, at least in the manner that is native to conventional understandings of moral personhood[13]. Ascribing notions of moral responsibility to [current] AI systems is a 'category mistake'[14].

2) Second, the context in which AI systems are being developed and deployed is overwhelmingly corporate in its complexion. It is, therefore, a context characterised by irresponsibility, at least insofar as that concept is understood here. Indeed, some forms of corporation – such as the Anglo-American style for-profit corporation – may be understood to be irresponsible *by design*[15]. As will be discussed below, in that form of corporation irresponsibility can be seen to emerge from the ways in which the corporate device distributes the risks and rewards associated with corporate conduct[16]. Rewards are largely distributed to those 'internal' to the corporation, whereas risks (including risks of harm) are largely distributed amongst those 'external' to the corporation[17]. What we can think of as the corporation's irresponsible distributive function has important implications for AI regulation, particularly AI regulation that is aimed at controlling for the risks associated with AI.

---

[10] A. Datta & T. Hartmann, *Commission withdraws AI liability directive after Vance attack on regulation*, «Euractiv», February 11, 2025, https://www.euractiv.com/section/tech/news/commission-withdraws-ai-liability-directive-after-vance-attack-on-regulation/.

[11] See for example T. Finn & A. Downie, *Agentic AI vs. generative AI,* «IBM Think», <https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai>.

[12] E.g. J. Gardner, "Relations of Responsibility" in *Crime, Punishment and Responsibility: The Jurisprudence of Antony Duff*, R. Cruft (ed) Oxford University Press, Oxford, 2011, p. 87; See P. Monti, *AI Enters Public Discourse: a Habermasian Assessment of the Moral Status of Large Language Models*, «Ethics & Politics», 61, 1, 2024, pp. 61-80.

[13] Moral personhood being broadly recognised as a pre-condition for ascriptions of moral responsibly. For a helpful treatment see F. Rudy-Hiller, *The Epistemic Condition for Moral Responsibility,* in «The Stanford Encyclopedia of Philosophy» (Winter 2022 ed), E.N. Zalta & U. Nodelman (eds.), https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/.

[14] N.H. Conradie and S.K. Nagel, *No Agent in the Machine: Being Responsible and Trustworthy about AI*, «Philosophy and Technology», 37, article 72, 2024, p. 5. Causal responsibility, of course, may be attributed to AI.

[15] E.g. P. Ireland, *Limited Liability, Shareholder Rights and the Problem of Corporate Irresponsibility*, «Cambridge Journal of Economics», 34, 5, 2010, pp. 837-856.

[16] K. Greenfield, *Saving the World with Corporate Law*, Boston College Law School Research Paper 130, 2007, pp. 10-11, http://dx.doi.org/10.2139/ssrn.978242 .

[17] *Ibidem*.

This piece explores aspects of the second claim. This treatment is intended to signal – but not exhaustively establish or explain – potential regulatory obstacles. In what follows I argue AI regulation must accommodate corporate irresponsibility, including the inherent, irresponsible distributive function exhibited by certain forms of the corporate device. Properly understood, AI regulation is a sub-set of corporate regulation. With few exceptions, the most appropriate point of departure in discussions around the design of AI regulation is not 'algorithms', but 'corporations', together with everything we have learned about corporate irresponsibility. The task of devising regulatory schemes for AI should be approached accordingly.

2. Some definitions

For the purposes of this discussion the terms 'regulation', 'regulatory schemes' and 'regulatory frameworks' means 'public control of a set of activities'[18]. Legislation (including subordinate legislation) would be the preeminent example of regulation under this definition[19]. The concept of 'AI regulation' should be read consistently with the above as formal rules devised and enforced by government concerning the development and/or use of AI technology. AI regulation need not take the form of horizontal or general regulation – it can operate in relation to discrete areas of the law. Existing regulation that does not specifically target AI systems but nonetheless captures such systems will not be AI regulation.

The concepts of regulation and AI regulation discussed in this piece are related to, but distinct from, the concept of the 'regulatory task', a concept discussed in Part 3.

'Artificial intelligence' means a "machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions…[they] vary in their levels of autonomy and adaptiveness"[20].

'Responsibility' is a word that houses many meanings[21] and here a number are invoked. In particular the discussion in this paper concerns notions of *legal* responsibility, both prospective and historic. Prospective legal responsibility exists if one has a legal duty to engage in certain conduct, or to avoid engaging in certain conduct[22]. Prospective legal responsibility has a forward looking, normative function[23]. Historic legal responsibility is backwards looking, is evaluative in nature, and concerns legal rules that ascribe legal accountability and/or liability for certain conduct, including sanctions[24]. In Part 6, the

[18] S. Chesterman, *We the Robots,* cit., p. 4.
[19] 'Soft law' as a form of private regulation falls outside the scope of 'regulation' in this piece.
[20] OECD, *Explanatory Memorandum on the Updated OECD Definition of an AI System*, «OECD AI Papers», 8, 2024, p. 4.
[21] For a helpful treatment of the emergence and trajectory of responsibility as a philosophical concept in western culture see generally R. McKeon, *The Development and the Significance of the Concept of Responsibility*, «Revue Internationale de Philosophie», 11, 39(1), 1957, pp. 3-32.
[22] P. Cane, *Responsibility in Law and Morality*, cit., pp. 31-32.
[23] Ivi, p. 57.
[24] *Ibidem*.

discussion also invokes a conception of moral responsibility taken from Williams, where responsibility '*represents the readiness to respond to a plurality of normative demands*'[25].

The term 'distributive function' is used here in two principal ways: first to refer to the distribution of risk and reward (between different sections of the community) that is produced by the operation of particular legislative or regulatory instruments[26]; and second, to refer to the distribution of risk and reward that is produced by the operation/design of corporate legal personality. As will be discussed in parts 5 and 6, particular forms of legal personality, such as corporate legal personality, are imbued with their own, internal programming around the distribution of risk and reward. As mentioned above, in the context of the Anglo-American style for-profit corporation, that distributive function operates predictably to internalise reward and externalise risk. The interaction between these two forms of distributive function (i.e. as it emerges from legislation/regulation, and as it emerges from corporate legal personality) is a central concern in the discussion below.

3. The regulatory task

Analytical engagement with regulatory phenomena is necessary built upon a particular understanding or conception of what can be thought of as the relevant 'regulatory task'[27]. The regulatory task that is the focus of this paper concerns governmental exercise of its regulatory function vis-à-vis emerging AI technologies in governance contexts characterised by representative government and market-based economic structures. By necessity, the regulatory task includes negotiation/debate informing the design and implementation of regulatory schemes. It may also include preliminary or 'placeholder' steps taken by governments that stop short of regulation, but which are nonetheless intended to influence the development and use of AI technology (codes of conduct, principles, white papers, etc).

Within such governance contexts, there appears to be two overarching concerns animating the regulatory task. These are respectively: 1) promoting innovation in AI; and 2) ensuring there are rules around the development and use of AI to mitigate against, or provide redress for, harms associated with the technology[28]. Given the nature of the technology and its possible applications, there is considerable tension between these two concerns. Indeed, at least insofar as innovation is understood to mean the development of novel technologies, then these two concerns may be regarded as largely antagonistic. While the scope of the regulatory task will evolve over time, in this early phase the task appears to involve identifying and establishing the legal landscape in which both of the above concerns may be pursued, albeit perhaps asymmetrically.

So conceived, and viewed from a high level of abstraction, the regulatory task is not limited to the striking of a balance between individual freedom and public interest (as is so often the case, particularly in contexts involving representative government). Rather, the

[25] G. Williams, *Responsibility as a Virtue*, «Ethical Theory and Moral Practice», 11, 4, 2008, pp. 455–470, p. 459.

[26] See for example P. Cane, *Responsibility in Law and Morality*, cit., p. 186-190, 219.

[27] See for example J. Black & R. Baldwin, *Really Responsive Risk-Based Regulation*, «Law & Policy», 32, 2, 2010, pp. 204-205.

[28] E.g. UK Government, *A Pro-Innovation Approach to AI Regulation*, (White Paper) 2023, <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>; EU AI Act, recitals 1 and 2.

regulatory task is one that Cane might describe as 'complex'[29] - it involves balancing *some* individual freedom against public interest, but also it involves attempting to balance *different, potentially competing public interest gains*. Here I am referring to the (largely anticipated) public interest associated with promised technological advancements in AI, as well as the obvious public interest that vests in preserving the safety of individuals, of communities, and the systems that both are reliant upon.

Understood in this way, the regulatory task is one that is strongly *distributive in nature*[30]. Which is to say, the task facing regulators is one in which they will take a hand in distributing the likely risks and likely benefits of AI technologies between different sections of the regulated community[31].

4. Distribution of risk and reward

In regulatory contexts involving market-based economies, the distribution of reward is one that is to a great extent shared with 'the market'. For example, the creation and provision of AI products/services, as well as the aggregated evaluations and choices of consumers with respect to those things, comprises a systemic crucible from which reward emerges. Likewise, the ultimate allocation of profits generated from AI market activity is ordinarily a matter of private ordering. Active governmental involvement in the distribution of reward will typically take the form of taxation, as well as legal protections for workers.

The fact that in market-based economies governments are content to share the task of distributing reward may help to explain why positive steps taken in pursuit of the regulatory task are largely concerned with distributions of risk. Notably, innovation friendly, harm mitigation as opposed to elimination[32] style regulatory postures (harm mitigation approaches) have already proven popular, and are perhaps the paradigmatic example of this early trend[33].

In a number of key jurisdictions, regulators appear to be framing the regulatory task with respect to AI as one that largely involves: first, identifying possible risks of harm associated with the development and use of AI (including economic, social and environment harms etc), as well as risks of harms associated with poorly designed regulation of such technology; second, deciding precisely how those identified risks of harm are to be distributed within a particular community; and third, designing and establishing legal responsibilities (both prospective and historic) aimed at producing the desired distribution of risk.

The above description highlights the central role played by considerations of responsibility in the design and implementation of AI regulation. The third, functionally distributive step described above involves regulators settling upon specific forms of *legal* responsibility so as to give expression to *moral/normative* allocations of responsibility agreed upon in the second step (with a particular focus there on questions around moral agency, such as accountability for harmful conduct

[29] P. Cane, *Responsibility in Law and Morality,* cit., p 251.
[30] Ivi, pp. 282-283.
[31] P. Cane, *Responsibility in Law and Morality*, cit., p. 186-190, 219; S. Chesterman, *We the Robots*, cit., 177-180.
[32] C. Coglianese, *The Law and Economics of Risk Regulation*, University of Pennsylvania, Institute for Law & Economics Research Paper No. 20-18, 2020, p. 9.
[33] Harm mitigation approaches are emerging in a number of jurisdictions, including the EU, Australia the UK, Canada and Singapore.

where feasible). Such a pattern of engagement with considerations of moral and legal responsibility is one that is deeply familiar in the regulatory context, especially in circumstances involving regulation intended to have a strongly distributive function.

5. Regulatory indifference on issues of legal personality

A final observation that can be made about the regulatory task is that regulators in various jurisdictions do not appear to be particularly concerned with sensitising performance of the regulatory task to the *identity* of the subjects of regulatory intervention. More particularly, regulators appear to be approaching the regulatory task armed with the tacit assumption that legal persons will respond to legal responsibilities created by AI regulation in a consistent manner.

Consider for example the EU's *AI Act*. This Act imposes various legal responsibilities upon 'providers', 'deployers', 'importers', 'product manufacturers' etc of AI technology[34]. Under Article 3 of the Act a 'provider', 'deployer', 'importer' etc could be a natural human being, or alternatively, various different forms of artificial/institutional legal personality such as for profit corporations, or public authorities. The default design of the law is to function in an undifferentiated way in relation to legal personality, absent some specified carve-out or exception. This approach arguably evidences a belief that the true levers of regulation are properly crafted legal responsibilities, where successful performance of the regulatory task involves getting the 'right' mix of prospective and historic legal responsibilities to target the 'right' legal persons.

This view of the regulatory task is not incorrect, but it is incomplete. Legal persons are *not* functionally neutral units. It is simply *not* the case that the imposition the same legal responsibilities across a range of different legal personalities can be expected to produce a desired behavioural result in a consistent manner. Different categories of legal personality – natural legal personality, corporate legal personality and bodies politic – typically exhibit quite distinct legal designs. Relevant differences include the purposes for which the grant of legal personality is made, the various legal capacities and obligations (including perhaps responsibilities) comprising the legal person, and finally, the existence of certain specified conditions imposed against a grant of legal personality, including conditions for use[35]. As might be expected, these differences in design produce differences in behaviour.

Critically for present purposes, a category of legal personality can come with its own internal or intrinsic programming with respect to *responsibility*, both moral and legal. This programming around responsibility can in turn create an intrinsic *distributive function* within a given category of legal person. If it is not taken into account, this intrinsic distributive function can obstruct the operation of external regulation where such regulation is intended to have a distributive effect. In Part 6 below I signal how the distributive function inherent in the Anglo-American style for-profit corporation – a distributive function which may be described as deeply *irresponsible* – may impede aspects of the regulatory task.

Focussing on this particular style of corporation is a necessary function of context. Unlike the industrial revolution in which unincorporated business models such as

[34] See for example *EU AI Act* article 2(1). These legal responsibilities are largely prospective in nature.
[35] M. Worthington & P. Spender, *Constructing Legal Personhood: Corporate Law's Legacy*, «Griffith Law Review», 30, 3, 2021, pp. 348-373.

partnerships played an important role[36], the AI revolution is predominately driven by large, even global, corporate interests. With the United States having emerged as an early leader in the development of AI, for-profit Anglo-American style corporation is playing an outsized role in shaping the nature and use of AI technologies.

6. Complicating the regulatory task – corporate legal personality and corporate irresponsibility

As noted above, understanding the intrinsic distributive function of the for-profit Anglo-American style corporation (from hereon in 'the corporation') is a matter of tracing the pattern of legal responsibilities – both prospective and historic – embedded within the design of the device. The configuration of these legal responsibilities in turn reflects determinations about the pattern of moral responsibility accorded to the corporation. With respect to the corporation, the most striking feature of each of these patterns of responsibility, legal and moral, is the extent to which they are oriented around a single concern – corporate financial self-interest.

Tracing first the pattern of legal responsibility, or irresponsibility, as it emerges from the corporation, there are four overlapping elements to consider.

First, corporate decision-making is subject to the operation of key prospective legal responsibilities, including the legal duty imposed upon directors that for convenience I will call the Best Interests Duty (BID). The precise content of the BID is jurisdiction dependent, however it is reasonable to describe it as requiring directors to act always in the best interests of the corporation, where corporate interests are ordinarily understood to equate to, or correlate with, the financial interests of the corporation's shareholders[37]. Critically, as shareholder interests are served by the creation of corporate profit, the BID operates to ensure that profit generation is prioritised above *all other considerations* in corporate decision-making. It is difficult to overstate the extent to which the BID trains directorial attention towards corporate financial self-interest. As noted by Greenfield:

> managers are held – or consider themselves held – to an obligation to take care of shareholders even when it hurts other stakeholders or society at large, and even when the benefits to those shareholders do not outweigh the costs to others[38].

It is even possible, though the question remains subject to debate, that under the BID 'managers have an obligation to violate *external* laws when necessary to meet their *internal* obligations to maximize returns to shareholders'[39]. As profit generation is aided by reducing the corporation's expenditures, the BID ensures that where possible, the corporation will *externalise its costs*, visiting them upon employees and their families, consumers, civic and political institutions, society at large (both present and future), other species, and the natural environmental systems upon which we all depend.

[36] P. Ireland, *Limited Liability, Shareholder Rights and Corporate Irresponsibility*, cit., p. 839.
[37] E.g. s172 *Companies Act 2006* (UK); s181 *Corporations Act 2001* (Cth); *Revlon Inc. v MacAndrews & Forbes Holdings Inc* 506 A.2d 173 (Del 1986).
[38] K Greenfield, *Saving the World with Corporate Law*, cit., p. 12.
[39] Ivi, p. 12, citing F.H. Easterbrook & D. R. Fischel, *Antitrust Suits by Targets of Tender Offers*, «Michigan Law Review», 80, 1982, pp. 1155-1178, p. 1177, n.57.

The operation of the BID is reinforced by the nature and function of shareholding in the corporation. This brings us to the remaining three elements for consideration, limited liability, separate legal personality and control rights.

Under the doctrine of limited liability, a shareholder's financial liability is limited to the cost of their investment in the corporation, i.e. the full price of their shares. Critically, a shareholders' personal assets are generally protected from claims relating to corporate liability. Moreover, separate corporate legal status ensures that it is the corporation, and not its shareholders, that has historic legal responsibility for corporate debts/liabilities. In the face of corporate collapse or corporate malfeasance, shareholders can lose the value of their investment, but nothing more. So shielded from legal responsibility, shareholders are far less likely to be concerned with the impacts, including negative impacts, of corporate conduct[40].

The nature and structure of corporate control rights compounds this phenomenon. As legal responsibility for day-to-day corporate management is ordinarily reserved for the board of directors, shareholders continue to derive an income stream irrespective of whether they have formed an understanding as to how corporate profits are generated[41]. Despite this fact, they also retain ultimate control over the corporation's activities – directors are accountable to shareholders, not least shareholders are (as a general proposition) legally responsible for the appointment and removal of directors. Moreover, shareholders are empowered to hold directors to their directorial obligations, including their obligations under the BID; an internal enforcement mechanism that helps to ensure that the corporation maintains its single-minded focus on profit generation[42].

How then, might we describe the pattern of moral responsibility that is revealed by the design of legal responsibility in the corporation? If the concept of responsibility is understood in the manner suggested by Williams, where responsibility 'represents the readiness to respond to a plurality of normative demands'[43], then the corporation is permitted to act, indeed required by law to act, in a manner that is profoundly *irresponsible*.

By design, the corporation is incapable of responding to a plurality of normative demands. Rather, the corporation responds to a *single* normative demand – advancing the financial interests of the corporation and, by extension, its shareholders. In practice, this response takes the form of an automatic, asymmetrical distributive function, whereby the corporation distributes the financial rewards of corporate conduct *internally*, and distributes the costs of corporate conduct *externally*. These externalised costs may include risks of harm, financial or otherwise, including grave risks. As it derives from the corporation's monist, self-referenced normative posture, the corporation's distributive function may be characterised as inherently irresponsible.

7. Corporate irresponsibility – implications for the regulatory task

If it is to be at all effective in realising intended distributions of benefits and risks associated with AI, regulation must be properly sensitised to the corporation's inherent and

[40] S. Bottomley, *The Responsible Shareholder*, Edward Elgar, Cheltenham 2021, pp. 175-180.
[41] P. Ireland, *Limited Liability, Shareholder Rights and Corporate Irresponsibility*, cit., p. 845.
[42] Ireland describes this mix of control rights with the 'no-obligation, no-responsibility, no-liability' nature of corporate shareholding as 'a recipe for irresponsibility'. Ivi, pp. 853-854.
[43] G. Williams, *Responsibility as a Virtue,* cit., p. 459.

irresponsible distributive function[44]. Faithful performance of the regulatory task requires conceiving of the task as a sub-set of corporate regulation, where the primary goals are anticipating and guarding against corporate irresponsibility.

Reframing the regulatory task in this way carries important implications for performance of the task. One of these implications is that harm mitigation approaches are a poor choice for AI regulation. This claim may be made in circumstances where the nature of the regulatory task is as described immediately above, and also as it is described in Part 3. The basis for this claim is that harm mitigation approaches are especially vulnerable to obstruction by corporate irresponsibility, at least insofar as they seek to distribute risk in a manner that differs to the corporation's own intrinsic distributive function. Given the monist design of the corporation's intrinsic distributive function, such differences are likely.

As noted in Part 3 above, the regulatory task tends to be directed towards two overarching concerns: 1) promoting innovation in AI; 2) and mitigating against, or providing redress for, AI harms. Further, as noted above, harm mitigation style regulation tends to establish these two concerns as antagonistic to one another; the less mitigation there is, the more innovation there is, etc. As innovation equals reward and mitigation equals cost, this antagonistic framing is easily hijacked by the corporation's intrinsic, irresponsible distributive function.

Consider for example, the *EU AI Act* (the Act). A combination function/risk-based instrument, in part the Act operates by classifying AI systems into different risk categories with attendant risk mitigation responsibilities. The most onerous risk mitigation obligations attach to systems categorised as 'High Risk' (HR). Interestingly, depending on the system, providers and deployers may play a role in determining whether a system is HR or not[45], potentially exempting themselves from the mitigation obligations that would otherwise attach to the system[46]. Where a system is classified as HR, providers play a role in shaping their mitigation obligations. For example, under article 9, providers of HR systems must create and maintain a 'risk management system', which involves, amongst other things, identifying risks that the provider must then attempt to mitigate. Deployers also have a role to play in shaping their obligations under the Act, including in identifying risks associated with their systems, and identifying 'serious incidents' relating to the use of their systems[47].

Even from this brief overview the flaws in this approach should be apparent. Risk mitigation approaches routinely require delegation to regulated entities, including corporations[48]. This involvement creates an opening for the corporation's intrinsic distributive function to displace the intended risk distributions of the regulatory instrument. Where providers or deployers are corporations, their intrinsic distributive function will incentivise under-classification of HR systems, understating of possible risks, and under-reporting of serious incidents, amongst other things. Doing so will alleviate costs and promote corporate financial gain[49]. The likelihood that corporations would seek to

[44] Cfr J. Black & R. Baldwin, *Really Responsive Risk-Based Regulation*, cit., pp. 188-197.
[45] See for example *EU AI Act* article 6(3).
[46] For an insightful critique of this 'loophole' see S. Wachter*, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What this Means for the European Union, the United States and Beyond*, «Yale Journal of Law & Technology», 26, 3, 2024, pp. 684-686.
[47] *Eu AI Act* article 26(5).
[48] J. Black & R. Baldwin, *Really Responsive Risk-Based Regulation*, cit., p. 201.
[49] Cfr Ivi, p. 202.

minimise or even evade mitigation obligations is enhanced by the obvious difficulties associated with detecting such behaviour, not least as 'conformity assessments' under the Act are undertaken by providers themselves, a mechanism aptly described by Wachter as 'a major legal loophole'[50]. Ultimately, the Act is insensitive to corporate irresponsibility and highly vulnerable to being obstructed by it as a consequence.

8. A possible solution

While it is beyond the scope of the present discussion to examine in detail how the problem of the corporation's inherently irresponsible distributive function might be overcome, it is worth mentioning in closing one possible solution. Arguably, the key to sensitising AI regulation lies in harmonising the different aims of such regulation. Regulators should view the twin concerns of promoting innovation in AI and mitigating potential AI related harms as being mutually reinforcing, rather than antagonistic, considerations. That is, regulators should approach the concept of innovation as *entailing* safety. Entangling the considerations in this way may help to confound the corporation's irresponsible distributive function. Disturbing the corporation's ability to treat innovation and risk mitigation as zero-sum propositions will necessarily impede the corporation's ability to externalise risks of harm, particularly where corporations are the developers of AI technology. Conceiving of innovation as entailing safety would train regulatory attention away from risk mitigation strategies, and toward approaches based on positive legal obligations to produce safe outcomes as a matter of fact. It will involve distributing the risks of AI harms back towards the developers and deployers of AI, building these costs back into the costs of development in particular. While they are hardly perfect models, product safety regulation and aviation regulation offer promising inspiration[51], as these areas tend to involve positive legal obligations to produce safe outcomes, rather than positive legal obligations aimed at merely reducing risk[52].

[50] S. Wachter, *Limitations and Loopholes*, cit., p. 692.

[51] Cf B. Judge et al, *When Code Isn't Law: Rethinking Regulation for Artificial Intelligence*, «Policy and Society», 44, 1, 2025, pp. 87-89. See also the *Australian Consumer Law*, and provisions such as section 54 which requires providers to ensure that goods and services are of an 'acceptable quality', including by being fit for purpose, acceptable in appearance, free from defects, safe and durable. Strictly speaking, section 54 operates as a guarantee, rather than a positive obligation, however the practical effect is quite similar, and hence it is interesting legislative artifact when viewed from a design inspiration perspective.

[52] This would likely involve adopting a 'zero risk approach' (eliminating potential harms) or an 'acceptable risk approach' (reducing risks of harm to acceptable thresholds): see C. Coglianese, *The Law and Economics of Risk Regulation,* cit., p. 9.

Ambivalenze digitali, tra potenzialità, miraggi e labirinti
Alla ricerca di un approccio etico "human-centered"[a]

Federico Reggio[*]

Abstract

Il presente contributo prende in esame due versanti particolarmente "topici" nel dibattito contemporaneo sulle tecnologie – Metaverso e IA – per esaminarne alcune criticità. Esse vengono lette come manifestazione dell'ambivalente rapporto tra essere umano e tecnica, cifra del moderno *homo faber* e tratto caratterizzante del suo erede contemporaneo, lo *homo tecnologicus*, che di questo "volto di Giano" sperimenta risvolti inquietanti e spersonalizzanti. Lo scritto si concentra in particolare su alcune vulnerabilità che espongono, nel mondo digitale, la persona umana, designando tre figure particolari di *digital discrimination*. In chiave costruttiva, si propone un cambiamento di prospettive, volto a immaginare una tecnologia (digitale) che sia pensata per essere e mantenersi "*human centered*": a tal fine si trae spunto dalla *restorative ethics*, quale etica attenta all'essere umano nella sua unicità, relazionalità e vulnerabilità.

Parole chiave: metaverso; IA; tecnologia human-centered; etica riparativa.

This contribution examines two particularly "topical" aspects of the contemporary debate on technology – the Metaverse and AI – in order to explore some of their critical issues. These are interpreted as manifestations of the ambivalent relationship between humans and technology, a hallmark of the modern *homo faber* and a defining feature of his contemporary heir, *homo tecnologicus*, who experiences the disturbing and depersonalising implications of this "Janus face". The paper focuses in particular on certain vulnerabilities to which humans are exposed in the digital world, identifying three specific forms of digital discrimination. In a constructive vein, it proposes a change of perspective, aimed at imagining a (digital) technology that is designed to be and remain *human-centred*: to this end, it draws inspiration from restorative ethics, an ethic that focuses on human beings in their uniqueness, relationality and vulnerability.

Keywords: metaverse; AI; human-centered technology; restorative ethics.

[a] Ricevuto il 02/07/2025 e pubblicato il 09/12/2025.
[*] Università degli Studi di Padova, e-mail: federico.reggio@unipd.it.

1. Una ambivalenza consolidata

La riflessione filosofica sembra da tempo evidenziare l'ambivalenza del rapporto tra essere umano e tecnica. Da un lato la *téchne* appare inscindibile rispetto al mondo umano, al punto da caratterizzarlo come dato tipico dell'umanità stessa, la quale, sin dagli albori, si è sviluppata anche in ragione della sua attitudine ad agire tecnicamente sul mondo, potenziando le proprie capacità. Nel contempo, però, la tecnica stessa ha consentito, in molti ambiti, di realizzare varie forme di 'sostituti' dell'essere umano, portando di volta in volta anche a ridefinire l'ambito dell'agire umano e, ancor più, lo spettro di quelle caratteristiche e attività che possono dirsi, propriamente, appannaggio dell'uomo. Si capisce, dunque, l'emergere di un profilo inquietante: come osserva, per esempio, Belardinelli, «scienza e tecnica hanno cessato di essere strumenti nelle mani dell'uomo e tendono a diventare sempre più fini a se stesse […]. Volevamo essere più liberi e ci ritroviamo inseriti in un'anonima processualità»[1].

L'attenzione verso i profili potenzialmente disumanizzanti della dimensione tecnica è, peraltro, un aspetto che si è manifestato in modo significativo nella riflessione del secolo scorso, soprattutto a partire dalla sua seconda metà. Il rischio che la tecnica sfugga al controllo dell'essere umano stesso, fino a rivolgersi contro a quest'ultimo, è stato da più parti evidenziato. Si pensi, emblematicamente, a quanto affermava Martin Buber: «l'uomo si lascia distanziare dalle sue stesse opere: così io esprimerei la peculiarità della crisi moderna. L'uomo non è più capace di signoreggiare il mondo che egli stesso ha fatto sorgere: questo mondo diviene più forte di lui, si libera di lui, gli sta innanzi nella sua elementare indipendenza, e l'uomo non conosce più la parola che abbia il potere di assoggettare il Golem che egli ha creato, e di renderlo inoffensivo»[2].

Sullo sfondo, dunque, troviamo un'ambivalenza: la tecnica come "estensione" dell'umano, e la tecnica come "sostituzione" dello stesso.

Non è questa, tuttavia, l'unica ambivalenza sottesa al rapporto tra tecnica e uomo. Martin Heidegger, per esempio, ne *La questione della tecnica*, evidenziava, appunto, un'ulteriore dualità: da un lato, il suo essere "un'attività dell'uomo" e, dall'altro, il suo essere mero «mezzo in vista di fini»[3]. Qui si cela, tuttavia, un vero e proprio elemento di crisi per l'uomo della tarda modernità, caratterizzato da una sostanziale perdita della capacità di pensiero teleologico: a esso, infatti, l'uomo moderno sembra aver progressivamente preferito un pensiero concentrato sul lato operazionale e calcolante, in cui il fine, se ancora sussiste, tende a essere appiattito all'utilità pragmatica[4]. Il risultato "veramente inquietante", per richiamare nuovamente Heidegger, «non è che il mondo si trasformi in un completo dominio della tecnica. Di gran lunga più inquietante è che l'uomo non è affatto preparato a questo radicale mutamento del mondo. Di gran lunga più inquietante è che non siamo ancora capaci di raggiungere, attraverso un pensiero pensante, un confronto adeguato con ciò che sta realmente emergendo nella nostra epoca»[5].

[1] S. Belardinelli, *Il gioco delle parti*, ATS, Roma 1996, p. 32.
[2] M. Buber, *Il problema dell'uomo*, Marietti, Genova 2004, p. 59.
[3] M. Heidegger, *La questione della tecnica*, in Id., *Saggi e discorsi*, Mursia, Milano 2007, p. 31.
[4] Si veda, sul punto, quanto evidenziato da R. Spaemann, *Persone. Sulla differenza tra 'qualcosa' e 'qualcuno'*, Laterza, Roma-Bari 2005.
[5] M. Heidegger, *L'abbandono*, Il Melangolo, Genova 1983, p. 31

Sebbene risalente, questa riflessione mi pare ancora di stringente attualità. In particolare, il termine "pensiero pensante" si rivela particolarmente provocatorio oggi, con lo sviluppo vertiginoso dell'Intelligenza Artificiale (IA). La contemporaneità è sfidata non solo da forme di tecnologia studiate per avocare a sé, sostituendole, funzioni "produttive" tipicamente umane: l'IA è costruita per imitare, coadiuvare e finanche sostituire processi cognitivi e creativi avanzati, spingendo in là il confine di ciò che già la rivoluzione informatica aveva avviato. Se già il computer ha sostituito da tempo funzioni prima demandate a risorse umane, offrendo peraltro strumenti di straordinario rilievo nell'ambito dello sviluppo complessivo delle società contemporanee, l'IA apre scenari di riflessione ulteriori, nel momento in cui essa mostra di appropriarsi di ambiti fino a ieri ritenuti intangibilmente appannaggio dell'umano stesso, sia esso un mero fruitore o un *designer* dello strumento tecnologico. Strumenti di IA capaci di realizzare scritti ed elaborazioni complesse, mostrando di realizzare "prodotti" un tempo pensabili solamente come prodotto della sola creatività umana, pongono di nuovo l'umano stesso davanti a uno specchio che lo interroga: anzitutto sui pericoli insiti nelle creazioni che ha realizzato ma soprattutto su cosa resti il suo *proprium*.

Sul punto, era stato quasi profetico Emanuele Severino: «già da tempo […] la dimensione umana è sempre più estromessa dal sistema di mezzi che sono coinvolti nel processo produttivo. Le "macchine", si dice, sostituiscono sempre di più l'uomo. E già da tempo si delinea la possibilità che l'Apparato stesso della tecno-scienza si configuri, nel suo insieme, come qualcosa che può essere interpretato come "Coscienza", "Sistema cosciente" o "Intelligenza Artificiale"»[6].

In questo senso, pur nella novità sollevata dall'emergere di queste nuove tecnologie, l'evoluzione del mondo digitale contemporaneo manifesta una ulteriore sfaccettatura di quel "volto di Giano" della tecnica, per il quale la *téchne* si presenta come capace di aumentare le "possibilità" operative dell'uomo stesso, e tuttavia, nel contempo, aumenta gli ambiti in cui l'umano può trovarsi "privato" di specificità, se non addirittura sottoposto alla potenza manipolativa della tecnica. Sono molti, infatti, gli ambiti in cui il contemporaneo *homo tecnologicus*, variante del moderno *homo faber*, percepisce il rischio di trovarsi *fabricatus*, oggettivato, e dunque, in fin dei conti, ben più che "sostituito": disumanizzato[7].

Non serve scomodare l'ampia e convergente produzione fantascientifica del XX secolo per notare come molti versanti dello sviluppo tecnologico contemporaneo rivelino, allo stesso tempo, potenzialità e pericoli[8]. Per richiamarci a nozioni e percezioni

[6] E. Severino, *Capitalismo senza futuro*, Rizzoli, Milano 2012, pp. 90-92.

[7] Cfr. G. Israel, *Il giardino dei noci: incubi postmoderni e tirannia della tecnoscienza*, Cuen, Napoli 1998; V. Possenti, *L'uomo postmoderno. Tecnica, religione e politica*, Marietti, Genova-Milano 2009. Sull'idea di *homo fabricatus* nella lettura di alcune proposte contemporanee particolarmente provocatorie rinvio ad A.L. Kiss, *"Zukunftsmensch" als Homo fabricatus. Bemerkungen über die futuristische Anthropologie von Jean Baudrillard und Peter Sloterdijk*, in «Sythesis Philosophica», vol. 61, n. 1, 2016, pp. 55-64. Cfr., altresì, F. Reggio, *Frontiere. Tre itinerari biogiuridici,* Primiceri, Padova 2018.

[8] Non è tuttavia irrilevante che dalla letteratura siano giunte raffigurazioni orientate a descrivere aspetti poi divenuti scenari reali o verosimili nelle successive evoluzioni storiche. Cfr., per una prima rassegna, E. Davenport, *Literature as Thought Experiment (On Aiding and Abetting the Muse)*, in «Philosophy of the Social Sciences», vol. 13, n. 3, 1983, pp. 279-306; P. Di Filippo, *A science-fiction fantasy*, in «Nature», 465, 2010, p. 1110, https://doi.org/10.1038/4651110a; J. Gunn, *Alternate Worlds: The Illustrated History of Science Fiction*, A&W, New York 1975; N.K. Hayles, *How we became posthuman: virtual bodies in cybernetics, literature, and informatics*, University of Chicago Press, Chicago 1999; D.A. Kirby, *The Devil in Our DNA: A Brief History of*

ampiamente diffuse, si pensi *in primis* alla capacità di "imbrigliare" e di produrre l'energia, fondamentale per lo sviluppo tecnico, industriale e sociale delle società moderne e contemporanee: essa ha mostrato anche un volto oscuro, consegnando all'umanità contemporanea incubi legati tanto al possibile uso offensivo-bellico di alte quantità di energia (con lo scenario di una possibile estinzione della vita sul pianeta, non contemplata anche nei peggiori scenari di guerra, anteriormente ai bombardamenti di Hiroshima e Nagasaki), quanto al risvolto ambientale degli strumenti destinati alla produzione energetica. Il tema è purtroppo di rinnovata attualità anche alla luce dei conflitti contemporanei, in cui l'incubo di uno scenario nucleare, così come la realtà di un elevatissimo inquinamento, si aggiungono alle già drammatiche considerazioni riguardanti le ingenti perdite di vite tra militari e civili.

Parimenti, la sempre più profonda conoscenza del *bios* umano, connessa alla sempre più ampia capacità di intervenire "tecnicamente" sulla vita biologica, mostra un susseguirsi continuo di grandi conseguimenti sul piano delle tecnologie biomediche, aprendo scenari diagnostici e terapeutici sino a poco fa inimmaginabili. Tuttavia, nello stesso tempo, questa conoscenza ha reso artificialmente manipolabili, potenzialmente disponibili, sfere e ambiti della vita ritenuti intangibili. Tutto ciò ha restituito l'inquietudine di un essere umano che può facilmente trovarsi ridotto a "oggetto" sperimentale, disponibile, posto nelle "mani" di chi ha la capacità tecnica di agire sulla vita stessa, intervenendo in modo anche radicale sulla sua origine, sulla sua fine, sulla qualità stessa sino al punto da aprire scenari di superamento del confine del concetto di "natura umana" (visibili, per esempio nel dibattito sul post-umanesimo e sul trans-umanesimo)[9]. Peraltro, una simile "invasività" non si limita alla sfera del "biologico" ma investe immediatamente anche quella del "politico" e del "giuridico", intaccando il mondo umano sotto una pluralità di distinti ma interconnessi profili[10].

Tornando allo sviluppo delle tecnologie informatiche, fondamentale per le contemporanee società complesse, esso ha aperto straordinari scenari atti ad ampliare la connettività, la rapidità e la raffinatezza di elaborazione di dati, potenziando capacità di analisi, di progettazione, dispiegando effetti di grande impatto su una pluralità di interconnessi livelli. Eppure, non sfuggono anche qui scenari inquietanti, se si pensa a un mondo sempre più concepito come un insieme di "dati", la cui circolazione e il cui utilizzo spesso sfuggono alla sfera di comprensione e controllo dei titolari (o delle sorgenti) di tali dati stessi[11]. Un mondo digitale, peraltro, la cui "virtualità" da un lato sottrae tempo e

Eugenics in Science Fiction Films, in «Literature and Medicine», vol. 26, n. 1, 2007, pp. 83-110; N. Miščević, *Political Thought Experiments from Plato to Rawl*s, in J.B. Brown, I. Frappier, L. Meynell (a cura di), *Thought Experiments in Philosophy, Science and the Arts*, Routledge, London 1991, pp 191-206; L.M. Silver, *Remaking Eden: cloning and beyond in a brave new world*; Avon Books, New York 1997. Cfr. altresì, su un piano più etico-filosofico, F. Palmer, *Literature and Moral Understanding: a Philosophical Essay on Ethics, Aesthetics, Education and Culture*, Clarendon Press, London 1992.

[9] Si veda sul punto, per esempio, il n. 3, 2019 di «Studium Ricerca», intitolato *Dopo di noi. Varianti e questioni del post-umano*, a cura di Antonio Allegra.

[10] Cfr., al tal riguardo, le profetiche considerazioni contenute in M. Foucault, *Nascita della biopolitica: corso al Collège de France (1978-1979)*, Feltrinelli, Milano 2019, e, più recentemente, anche V. Possenti, *La rivoluzione biopolitica. La fatale alleanza tra materialismo e tecnica*, Lindau, Torino 2013, pp. 217-219.

[11] Non deve sfuggire, a riguardo dei dati, il riduzionismo insito nel fenomeno della cosiddetta "datificazione", per cui rinvio, per un'analisi critica, a C. Sarra*, Il mondo-dato. Saggi su datificazione e diritto*, Cleup, Padova 2022. Cfr., altresì, M. Durante, U. Pagallo (a cura di), *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società*, Mimesis, Milano 2022. Sulla tendenza nichilistica di questo fenomeno,

risorse alla vita reale (si pensi alle giovani generazioni che vivono "attaccate" al loro smartphone) e dall'altro lato dispiega effetti anche potenzialmente devastanti sulla qualità della vita stessa (sotto molteplici profili: dell'informazione[12]; delle relazioni[13]; della salute[14]).

La tecnica, dunque, sembra, in molti ambiti, sottrarre l'uomo da un confronto con se stesso e, nel contempo, offrirgli l'occasione di porsi allo specchio, laddove si sia capaci di non soffocare la possibile inquietudine che ciò comporti, magari soffocandola in mille anestetizzanti distrazioni. In questo bivio, dunque, resta possibile accettare la sfida di interrogarsi su cosa caratterizzi l'umano come tale e su come eventualmente difendere questa sua prerogativa dal pericolo che venga negata, o comunque indebitamente compressa. Sullo sfondo, peraltro, emerge l'ancor più radicale quesito sulla sensatezza di considerare "l'umano" meritevole di tutela in quanto tale. Un tema, invero, per nulla scontato, visto che la fondatezza di una simile domanda è stata ampiamente discussa, per esempio, in ambito bioetico. «Perché è moralmente sbagliato sopprimere una vita umana?» (…) «cosa c'è di così speciale che una vita sia umana?»: questa domanda non se la pone un oscuro scienziato al soldo di un regime totalitario, bensì Peter Singer, giurista australiano di origine ebree, per anni docente a Princeton e pioniere del movimento per i diritti degli animali ed esponente di spicco della bioetica *pro choice*[15].

Questo ci conduce a un quesito fondamentale nel contesto della nostra riflessione: una volta evidenziato che la relazione tra essere umano e tecnica presenta *profonde e molteplici ambivalenze*, è necessario comprendere se e in che misura queste ultime *possano dipendere dalla prospettiva con cui l'essere umano si approccia alla tecnica stessa*.

La questione, insomma, intreccia un quesito etico con uno antropologico.

2. Solo homo faber?

Non manca chi osserva, invero, che vi sarebbe una reciproca implicazione tra il "fare" della tecnica e il "farsi" (e il "comprendersi") dell'essere umano. Scrive, ad esempio, Anna

immerso nelle moderne "infocrazie", si vedano anche le brevi e incisive considerazioni di Byung-Chul Han, per cui cfr. B. Han, *Infocrazia. Le nostre vite manipolate dalla rete*, Einaudi, Torino 2023. Il tema tecnologico, peraltro, si rinsalda fortemente con quello bioetico, come osservato, per esempio, in L. Palazzani, *Dalla bio-etica alla tecno-etica: nuove sfide al diritto*, Giappichelli, Torino 2017; S. Amato, *Biodiritto 4.0. Intelligenza artificiale e nuove tecnologie*, Giappichelli, Torino 2020; C. Sartea, *Ecotecnologia. Sfide etico-giuridiche della civiltà tecnologica*, Giappichelli, Torino 2024.

[12] Si veda, emblematicamente B. Chul-Han, *Infocrazia. Le nostre vite manipolate dalla rete*, cit., e, per quanto attiene al tema 'informazione e verità', F. D'Agostini, M. Ferrera, *La verità al potere*, Einaudi, Torino 2019; D. Butturini, *Il diritto alla ricerca della verità: profi li costituzionali tra partecipazione democratica e lotta alla manipolazione*, in F. Reggio (a cura di), *Honeste Vivere. Percorsi filosofici per l'etica pubblica*, FrancoAngeli, Milano 2025, pp. 173-210.

[13] Mi sia consentito citare, come esempio significativo, un fenomeno molto presente sui social negli ultimi anni, emblematico di un deterioramento delle relazioni in senso polarizzante, ossia la *cancel culture*. Cfr. F. Reggio, *Cancel culture e questioni di etica pubblica. Memoria, celebrazione ed esecrazione pubblica nelle democrazie contemporanee*, in F. Reggio (a cura di), *Honeste Vivere. Percorsi filosofici per l'etica pubblica*, cit., pp. 133-172.

[14] Su quest'ultimo punto basti esaminare il documento della World Health Organization, *Addressing Health Issues in the Digital World*: https://www.who.int/europe/publications/i/item/WHO-EURO-2024-10917-50689-76724.

[15] P. Singer, *Ripensare la vita. La vecchia morale non serve più* (1994), tr. it. di S. Rini, Il Saggiatore, Milano 1996, pp.114-115. Per Singer, rivendicare una dignità peculiare all'essere umano è una forma di 'specismo' ossia un atteggiamento che indebitamente avanza una sorta di supremazia della specie umana su altre specie viventi e senzienti, similmente a quanto il razzismo fa discriminando tra diverse razze.

Pintore: «l'onnipresenza della tecnica, e del pensiero scientifico che ne ha reso possibile lo sviluppo oggi tumultuoso, non è il prodotto di una certa visione dell'uomo a cui sia possibile opporne una differente – perché essa in realtà è qualcosa che precede e rende possibile il prodursi di qualunque visione dell'uomo»[16]. In quest'ottica, dunque, nella relazione tra uomo e tecnica non sarebbe rilevante la prospettiva *dell'*umano e *sull'*umano che si assume.

L'uomo non potrebbe, dunque, essere *nient'altro che homo faber* (e, aggiungiamo, *tecnologicus*), perché un'antropologia che prescindesse dal «fatto che l'uomo abbia le mani, i sensi e dunque la capacità oltre che l'istinto di conoscere e controllare la natura» prescinderebbe dalla *natura* stessa dell'uomo: «negare l'essenzialità del rapporto manipolatorio dell'uomo con il mondo non è semplicemente proporre un'antropologia alternativa, perché in realtà nessuna antropologia può prescindere da questo presupposto, anche se lo ignora, lo nega o lo combatte»[17]. Appare in effetti difficile osservare e capire l'essere umano – e tanto meno, come si è detto, l'uomo occidentale – se si astrae dal suo rapportarsi al mondo circostante anche in termini di manipolazione e trasformazione. Da che ha coscienza della sua presenza nel mondo, l'essere umano si intreccia con il mondo stesso cercando di utilizzarne risorse ed energie per la propria sopravvivenza e per migliorare le proprie condizioni di vita, nonché per affermarsi[18]. Il prometeico, dunque, è, almeno in certa misura, parte dell'uomo, il quale si è manifestato, agendo sul mondo, manipolando e trasformando. Si tratterebbe, peraltro, di una consapevolezza antica, se pensiamo al fatto che nell'*Antigone* Sofocle usa il termine *deinós* (δεινός) riferendosi alla stupefacente capacità degli uomini di essere terribili e allo stesso tempo meravigliosi: abili – pare lecito pensare – nel costruire e nel distruggere.

Ora, che nell'uomo risieda *anche* l'essere *faber*, appare dunque un'affermazione dotata di evidente peso argomentativo. Non è tuttavia univocamente accettata l'idea che l'uomo non possa essere altro che *homo faber*. Ciò porterebbe a porre in ombra quanto già Hannah Arendt emblematicamente rilevava, ossia su come il modello antropologico dell'*homo faber* rappresenti un tratto specifico e predominante della mentalità moderna in Occidente. La studiosa evidenzia, peraltro, che l'affermarsi di questa figura è accompagnata da uno spostamento della "messa a fuoco" del pensiero e, di conseguenza, dell'agire pratico: essa comporta, dunque, anche una parziale ridefinizione di ciò che è specificamente umano. All'esito di questa "svolta antropologica" cambiano percezioni, priorità, domande fondamentali: per esempio, ciò che importa massimamente non più il *che cosa* (il *ti esti* dell'interrogare socratico) o il *perché* (inteso in senso causale e, ancor più, in senso teleologico), bensì il *come*. Un "come" orientato allo scoprire meccanismi, in vista del loro riprodurli e dominarli tecnicamente[19]. Il punto nodale risiede proprio nell'attitudine verso

[16] A. Pintore, *Il desiderio dei diritti*, in «Rivista di Filosofia del Diritto», n. 2, 2017, p. 248.

[17] Ivi, p. 247.

[18] Nella psicologia sociale di George Herbert Mead, per esempio, solo attraverso l'interrelazione "manipolativa" con il "mondo" l'uomo prenderebbe coscienza di sé come "nel mondo". Cfr., a titolo esemplificativo, G.H. Mead, *The Individual and the Social Self: Unpublished Essays*, Chicago University Press, Chicago 1982.

[19] Ciò è visibile negli atteggiamenti tipici dell'*homo faber*: la sua strumentalizzazione del mondo, la sua fiducia negli strumenti e nella produttività del costruttore di oggetti di artificiali, nella portata omnicomprensiva della categoria mezzi-fine, la sua convinzione che ogni problema possa essere risolto e ogni motivazione umana ridotta al principio dell'utilità; la sua sovranità, che considera come *un immenso tessuto da cui possiamo ritagliare ciò che vogliamo*, la sua equiparazione di intelligenza ed ingegnosità, cioè il suo disprezzo per ogni

l'agire che caratterizza il presupposto antropologico *de quo*: l'agire dell'*homo faber* non esaurisce la *vita activa*, né costituisce l'unica formula per l'agire umano. Anzi, esso, nei suoi tratti fondamentali, rappresenta, nel contempo, una riduzione e una "forzatura" delle potenzialità umane, al punto da comprimere o persino trascurare altri profili antropologici che, soprattutto in epoca premoderna, non riducevano l'essenza dell'umano alla sua capacità tecnica (valorizzando, ad esempio, la dimensione *contemplativa*)[20].

In questo senso, le riflessioni pocanzi proposte aiutano anche a cogliere un aspetto del presente, così fortemente radicato in premesse (e fallimenti) dell'epoca moderna. La parabola di quest'ultima, infatti, se letta diacronicamente, sembra tracciare un percorso di progressivo restringimento del campo del pensiero e dell'agire, in cui si assiste ad una sempre maggiore svalutazione delle dimensioni del "contemplare", dell'interrogare, dell'agire come *prattein*, in funzione di un fare produttivo sempre più invasivo (visibile, tra l'altro, in una cultura diffusa che, oggi, accorda un privilegio alla tecnica sulla scienza, oltre che sulla filosofia)[21].

Dunque, anche accettando il fatto che possa dirsi strutturale, per l'uomo, essere capace di tecnica, e quindi di porsi in un rapporto manipolatorio e trasformativo con il mondo, ciò non significa necessariamente anche che egli *sia condizionato* a concepire il mondo stesso meramente come oggetto di dominio, manipolazione e trasformazione, né che ciò sia l'unica opzione possibile per leggere e dirigere le "mani" umane. Il vero problema, sia etico che filosofico, intorno all'umano, non riguarda il fatto di *avere le mani* bensì *l'uso che se ne fa*. Si ripresenta, quindi, prepotente e urgente, la domanda sul *fine* e sul *modo* attraverso i quali una certa capacità viene diretta, e questo tema spalanca orizzonti di riflessione anche sulla *visione del mondo* in forza della quale si concepisce e/o giustifica ciò che tali mani possono toccare, ed eventualmente entro quali limiti.

Se accettiamo che intorno all'*homo faber*, in passato "parte ausiliaria dell'*homo sapiens*", si è determinata, in epoca moderna, una "svolta antropologica", dobbiamo concludere che questa ridefinizione dell'uomo non esaurisce le possibilità e i modi di pensare l'umano[22].

Ritornare alla domanda antropologica si presenta, dunque, con particolare urgenza, perché il riduzionismo attivato dall'*homo faber* moderno è ben visibile oggi nei suoi esiti più espliciti e compiuti: un tempo era la tecnica ad essere «al servizio dell'esistenza umana, guidata dalla saggezza. Ora invece la ragione strumentale prende il predominio […] e conduce ad una sorta di atrofizzazione dell'identità dell'uomo»[23]. Di questa *eterogenesi dei fini*, per cui la tecnica, che doveva *servire*, finisce, come si è detto, per *asservire* lo stesso *homo faber*,

---

pensiero che non possa essere considerato come orientato alla fabbricazione di oggetti artificiali o di strumenti utili in tal senso; «infine la sua identificazione acritica della fabbricazione con l'azione» (H. Arendt, *Vita activa. La condizione umana*, Bompiani, Milano 2000, p. 227. Il corsivo è la citazione che Arendt stessa propone di un'espressione contenuta in H. Bergson, *L'Évolution créatrice (1907)* - per cui cfr. H. Bergson, *L'evoluzione creatrice*, Athena, Milano 1925).

[20] Per una prima rassegna in merito al recupero della dimensione *prattein* nell'etica della seconda metà del XX secolo, cfr. A. Vendemiati, *Universalismo e relativismo nell'etica contemporanea*, Marietti, Genova 2007, p. 149 e ss.

[21] Questo processo può essere letto anche alla luce del fenomeno della secolarizzazione: cfr., per una prima rassegna, L. Palazzani (a cura di), *Filosofia del diritto e secolarizzazione. Profili giuridici ed etici*, Studium, Roma 2011. In particolare, derivo l'idea di un "restringimento" del campo del pensiero e dell'azione dalla lettura che della secolarizzazione viene offerta in F. Cavalla, *All'origine del diritto al tramonto della legge*, Jovene, Napoli 2011, pp. 161-205.

[22] A. Vendemiati, *Universalismo e relativismo*, cit., p. 31.

[23] *Ibidem*.

l'Occidente sembra aver progressivamente iniziato a prendere coscienza nel crinale della post-modernità[24].

L'umanità contemporanea – soprattutto quando vuole riflettere sulle molte inquietudini che la avvolgono – raccoglie infatti dall'uomo moderno un'eredità travagliata, densa di crepe, timori e disillusioni, in cui il *gap* tra ciò che l'uomo "può fare" e ciò che egli effettivamente "sa" non restituisce "magnifiche sorti e progressive", ma anche paure e la percezione di un perdurante stato di crisi che si riverbera su vari aspetti della sua vita, personale e politica[25].

La riflessione filosofica sul rapporto tra tecnica e mondo umano apre scenari di pensiero sul modo in cui può essere concepita e attuata la relazione tra essere umano e mondo, nonché tra esseri umani: ciò, peraltro, conduce anche alla domanda – dal netto tenore etico – sulla responsabilità che l'essere umano stesso ha circa il modo in cui concepisce e attua tale relazione[26].

Come ho già avuto modo di evidenziare, a mio avviso occorre chiedersi se sia possibile *liberare l'umano* dal restare intrappolato nella figura dell'*homo faber*, il quale, nel concepire il mondo quale "ammasso di materiali" disponibile al suo agire manipolativo e trasformatore, ha progressivamente identificato il limite – proprio e del proprio agire – con una costrizione estrinseca che va rimossa. Così facendo, tuttavia, egli ha finito col perdere di vista il limite *tout-court*. Preoccupato di rimuovere "ostacoli al proprio agire" egli si è trovato a privarsi anche di forme di autentica "protezione" verso un uso incontrollato dell'agire proprio e altrui. Probabilmente si può spiegare così il motivo per cui si è arrivati, nel nostro presente, a quello che Claudio Sartea chiama icasticamente un «assedio tecnologico della persona»[27].

Certo, come è stato osservato, «le tecnologie e le biotecnologie possono avere usi buoni e cattivi – di solito il problema è che hanno usi buoni *e* cattivi insieme – da valutare caso per caso in modo equanime, cercando di non farsi paralizzare dalla paura dell'ignoto e del futuro»[28]. Questo è un monito di fondamentale importanza, soprattutto quando si considerino criticamente alcune "frontiere" della tecnologia contemporanea, come nel caso

---

[24] Non sono invero mancate voci "fuori dal coro" come quella di Giambattista Vico, che – già nel momento più propulsivo dell'affermarsi della mentalità moderna – elevava critiche circostanziate a quell'alleanza fra razionalismo, individualismo e utilitarismo che andava caratterizzando, con sempre maggiore forza, il pensiero filosofico, politico e giuridico già pre-illuminista, assumendo ben precise opzioni di pensiero criticabili sia nei presupposti che negli esiti. Sul rapporto fra Vico e la modernità cfr. M. Lilla, *G. B. Vico, the Making of an anti-modern*, Harvard University Press, Cambridge (MA) 1993; E. Voegelin, *la Scienza Nuova nella storia del pensiero politico,* Guida, Napoli 1996; R. Caporali, *Vico: quale modernità*?, in «Rivista di Filosofia», 1, 1996, pp. 357-378; G. Cacciatore, S. Caianiello, *Vico anti-moderno*, in «Bollettino del Centro di Studi Vichiani», voll. XXVI-XXVII, 1996/1997, pp. 205-218; A. Battistini, *Vico tra antichi e moderni*, Il Mulino, Bologna 2004; F. Reggio, *Il paradigma scartato. Saggio sulla filosofia del diritto di Giambattista Vico*, Primiceri, Padova 2021.

[25] Non è possibile dar conto della vastissima letteratura sul punto, ma mi sia concesso riferirmi in particolare a un'opera cui sono debitore per i preziosi spunti di pensiero: cfr. B. Montanari (a cura di), *La possibilità impazzita. Esodo dalla Modernità*, Giappichelli, Torino 2005. Non va dimenticato, poi, quanto gli anni '20 si siano aperti con una profonda crisi, anche intorno alla pandemia del Covid-19, che ha sollevato non pochi interrogativi di carattere bioetico, politico, giuridico, sociale. Cfr., per una prima disamina, F. Reggio, *La krisis del Coronavirus. Una sfida inattesa per l'essere umano e le società contemporanee. Considerazioni filosofico-giuridiche,* in «Calumet – intercultural law and humanities review», 1, 2020, pp. 118-142.

[26] Inevitabile il riferimento a H. Jonas, *Il principio responsabilità. Un'etica per la civiltà tecnologica*, Einaudi, Torino 2002.

[27] C. Sartea, *Ecotecnologia*, cit., p. 49.

[28] A. Pintore, *Il desiderio dei diritti*, cit., p. 248.

delle riflessioni nell'alveo delle quali si colloca il presente scritto, il quale si affaccia su scenari e problematiche legate agli sviluppi vertiginosi dell'informatica e della digitalizzazione.

Essere consapevoli della ambivalenza, della poli-potenzialità della tecnica, ci rammenta di non cadere nell'errore di demonizzare (scienza e) tecnica in quanto tali e, soprattutto, ci ammonisce circa il rischio opposto, ossia quello di un entusiasmo acritico: ne consegue che gli usi delle tecnologie vanno sottoposti a vaglio critico e contestuale, e non assunti a-criticamente come buoni o cattivi in quanto tali, cadendo negli opposti estremi del *tecno-entusiasmo* e della *tecno-fobia*.

Né vale la considerazione per la quale è già più volte avvenuto nella storia che l'emergere di una nuova tecnologia abbia profondamente scosso il mondo umano nella sua percezione di sé, nella sua organizzazione produttiva ed economica, nei suoi modelli di *governance*: la rivoluzione digitale contemporanea non sarebbe che uno di tanti "scatti in avanti" avvenuti nella storia umana, dopo la stampa, l'industrializzazione, l'informatizzazione, per cui la presente fase va letta come un momento di transito che avrà necessariamente un impatto forte su schemi consolidati, seguito da varie scosse di assestamento e poi da una normalizzazione. Che vi sia della ricorsività nelle "rivoluzioni tecnologiche" e nel loro impatto sulla vita umana a vario livello non v'è dubbio: che questa ricorsività ci trovi perciò stesso preparati ad affrontare la contemporaneità è però tutt'altro che scontato, così come non si può dare per garantita la capacità umana di "assestarsi" automaticamente sull'all'esito di un cambiamento, quasi entro un "mito evoluzionistico"[29].

Non bisogna dimenticare, a questo riguardo, il monito con cui Vico, al termine della *Scienza Nuova*, rammentava della possibilità di novelle barbarie capaci di gettare il mondo umano nella catastrofe; da essa l'umanità, Vico stesso lo ammette, può nascere rinnovata, ma spesso ad altissimo prezzo, e in ogni caso ciò è tutt'altro che scontato[30]. Ciò che però più conta, a mio avviso, nel monito vichiano, risiede nel fattore che per l'autore conduce al pericolo di queste novelle barbarie: ossia la perdita di vista del limite che costitutivamente caratterizza l'essere umano. Una perdita che può avvenire per effetto di una ragione "instupidita", o anche per l'apparente opposto, ossia per una "riflessiva malizia" che conduca a un eccesso di fiducia nelle capacità stesse dell'uomo di conoscere e dominare il mondo che lo circonda[31].

Questa esigenza di discernimento, tuttavia, costituisce, secondo me, uno scenario che offre una rimarchevole opportunità: essa spinge, infatti, a muoversi "oltre" l'*homo faber,* al suo pragmatico utilitarismo e ai suoi sogni di dominio. Permane, anche oggi, la possibilità di pensare l'essere umano in modo diverso dal suo "avere le mani", riscoprendo, per

[29] Eric Voegelin vede in questo una tipica caratteristica della modernità occidentale, ossia la *hybris* dell'autosalvazione (che non è altro che un alimentare nuovamente il mito della tecnica come strumento salvifico e come fine, a ben pensare). E. Voegelin, *la Scienza Nuova nella storia del pensiero politico*, Guida, Napoli 1996.

[30] Cfr. *Sn44, Conchiusione,* in G.B Vico, *Opere*, Mondadori, Milano 2000, p. 1109. Cfr., per una rilettura in merito alle "rinnovate barbarie", L. Bellofiore, *Morale e Storia in G. B. Vico*, Giuffre, Milano 1972; U. Galeazzi, *Ermeneutica e Storia in Vico. Morale, diritto e societa nella 'Scienza Nuova'*, Japadre, Roma-L'Aquila 1993; F. Reggio, *Il Paradigma scartato. Saggio sulla filosofia giuridica di Giambattista Vico*, cit.

[31] Come osserva Maria Donzelli, Vico «sembra avere presentito la dialettica dell'Illuminismo e i rischi di una forma di razionalismo assoluto e autoreferenziale. Egli ha chiara l'idea che la cultura, frutto del progresso tecnologico e scientifico, può condurre in una nuova condizione di barbarie» (M. Donzelli, *L'età dei barbari. Giambattista Vico e il nostro tempo*, Donzelli, Roma 2019, p. 100).

esempio, l'attualità e l'urgenza di ritrovare l'*homo dialogicus*, il quale, pur riconoscendo "l'importanza di valori esaltati nell'età moderna", li assume "in maniera profondamente diversa da come erano intesi dall'*homo faber*", in ragione di una diversa attitudine, in particolare nella relazione con i propri simili e con il mondo circostante: «l'incontro con l'altro lo arricchisce sia per quanto gli comunica sia per l'opportunità che gli offre di prendere coscienza dei suoi limiti»[32]. Questo, però, presuppone di elevarsi oltre lo scenario di un mondo composto da oggetti, dati ed elementi disponibili e manipolabili, aprendo la domanda sul senso e sulla dignità intrinseci a ciò che viene a trovarsi "nel mondo".

Mi sia tuttavia consentito nuovamente di ritornare a una riflessione che ho già espresso altrove, e della cui fondatezza e urgenza sono sempre più convinto: prima ancora di (ri)abilitare l'*homo dialogicus*, è anzitutto necessario ritrovare l'*homo sapiens*, ossia l'uomo capace e desideroso di pensare a se stesso nella consapevolezza di non poter uscire da questa domanda antropologica con un'attitudine (auto)oggettivante, e quindi salvaguardando, insieme al proprio mistero, la propria intrinseca dignità di soggetto[33].

3. Sapienza, indigenza, limite

Tra i molteplici modi in cui possiamo immaginare l'idea di sapienza, sottesa all'aggettivo *sapiens*, mi piace ricordare una definizione proposta da Roland Barthes, a cui spesso mi richiamo, quasi come si fa con un *mantra*: «nessun potere, un po' di sapere, un po' di intelligenza, e quanto più sapore possibile»[34]. La sapienza non è mera erudizione, non è strumento di potere, non è pura intelligenza: essa è intrinsecamente connessa con la coscienza del limite che, sola, consente di *dare sapore* all'avventura umana nel tempo e nello spazio[35]. Qui si rivela una preziosa ambivalenza del concetto stesso di limite, il quale non agisce puramente come fattore contenitivo bensì anche come elemento propulsivo, dal momento che, mostrando all'uomo la sua costitutiva indigenza di verità, lo apre tanto alla relazione con il proprio simile – rivelandolo irriducibile a *res* – quanto alla ricerca, in dialogo con quest'ultimo[36].

---

[32] F. Zanuso, *Conflitto e controllo sociale nel pensiero politico-giuridico moderno*, Cleup, Padova 1993, p. 20.

[33] Come evidenziato, emblematicamente, in S. Fuselli, *La lanterna di Diogene: alla ricerca dell'uomo negli esperimenti di ibridazione*, in F. Zanuso (a cura di), *Il Filo delle Parche. Opinioni comuni e valori condivisi nel dibattito biogiuridico*, FrancoAngeli, Milano 2009, in particolare pp. 100-104.

[34] R. Barthes*, Roland Barthes au Collège de France*, IMEC, Saint-Germain la Blanche-Herbe 2002, p. 12.

[35] Mi piace rammentare un *locus* letterario interessante a questo riguardo, perché pone a confronto due uomini intelligenti e potenti, ma diversamente sapienti (o meglio, l'uno sapiente, l'altro no): «Lo guardai – disse Gandalf – e vidi che le sue vesti non erano bianche come mi era parso, bensì tessute di tutti i colori, che quando si muoveva scintillavano e cambiavan tinta, abbagliando quasi la vista. "Preferivo il bianco", dissi. "Bianco!", sogghignò. "Serve come base. Il tessuto bianco può essere tinto. La pagina bianca ricoperta di scrittura, e la luce bianca decomposta". "Nel qual caso non sarà più bianca", dissi, "e colui che rompe un oggetto per scoprire cos'è, ha abbandonato il sentiero della saggezza"» (J.R.R. Tolkien, *Il signore degli Anelli*, BUR, Milano 1982, p. 327). Il dialogo, come è stato ricordato, avviene fra Gandalf e Saruman, un mago un tempo saggio, poi «inebriato da un miraggio faustiano di conoscenza e potere personale», il quale non a caso si avvale di macchine e fabbriche, di tecniche anche vòlte all'ibridazione fra razze e individui, e il cui disegno di potere si accompagna alla devastazione della natura e all'assoggettamento di popoli più inermi e benevoli. Cfr., sul punto, M. Manzin, *La natura (del potere) ama nascondersi*, in F. Cavalla (a cura di), *Cultura moderna e interpretazione classica*, Cedam, Padova 1997, pp. 85-112.

[36] Sul punto, richiamo le pregnanti riflessioni che Francesco Cavalla propone in margine ad Agostino, relativamente alla *sapientia* come presa di coscienza dell'insostenibilità dello scetticismo e del razionalismo dogmatico, e dell'esigenza di porsi in un'incessante ricerca di una verità che, innegabile razionalmente, non

La sapienza, dunque, abilita e nel contempo "obbliga" l'uomo (sia pur non fattualmente, perché non opera come costrizione) a interrogarsi sui confini del proprio agire, assumendosi la responsabilità – prima ancora dei limiti concretamente "posti" – della propria stessa finitudine[37]. Se la finitezza, infatti, è un aspetto condiviso da ogni essere umano, quale cifra dell'*ex-sistere* di ciascuno, essa «impone di manifestarsi come unico ed irripetibile all'interno di un orizzonte comune, alla luce del quale si è chiamati a rispondere e ancor prima a formulare domande. Per esistere bisogna riconoscere la dialetticità della compresenza del diverso e del comune; e testimoniarla con l'attiva responsabilità»[38].

Questa consapevolezza etica deve potersi proiettare anche sulla concretezza, aprendo quesiti che investono il modo in cui ci si approccia alle tecnologie e – per tornare al tema di questo scritto – al mondo digitale: ciò riguarda tanto gli utenti quanto i *designers*, i quali, nel progettare e nel promuovere strumenti digitali, possono fortemente condizionare la vita degli utenti dei loro prodotti, con evidenti implicazioni etiche e sociali.

Il "volto di Giano" della tecnica, pertanto, non deve intendersi alla stregua di una "esimente" di responsabilità – quasi si dovesse accettare come ineluttabile il rischio che i prodotti della tecnica si ritorcano contro i loro fruitori, a partire da alcune categorie particolarmente vulnerabili.

La responsabilità attiva che evochiamo in chiusa di queste pagine dal taglio più etico-generale invoca una riflessione rafforzata sul modo in cui una certa tecnologia viene progettata, veicolata, resa disponibile, nella consapevolezza che tale impatto può avere implicazioni dotate di risvolti alquanto problematici, se non disumanizzanti. Il *designer*, dunque, *non può pretendere una sua neutralità etica*, nascondendosi dietro alla presunta neutralità della tecnica stessa: anzi, proprio il volto ambivalente della tecnica è ciò che maggiormente richiama l'attenzione su chi può contribuire a delinearne e a indirizzarne i prodotti, delineando in capo a costoro una rafforzata responsabilità: etica e sociale, prima ancora che giuridica. Allo stesso tempo, anche il fruitore non può fare dell'immersione tecnologica un'esimente dalla riflessione sulle possibilità, sui limiti e sulle responsabilità che caratterizzano il suo agire. Per quanto "caratterizzato" da una tecnologia onnipresente e che influisce su conoscenze, comportamenti, e mentalità, l'uomo non deve dimenticare che la tecnica fornisce strumenti, e quindi mezzi. *«Eadem turpia aut honesta fieri possunt»* - ricordava Seneca nell'Epistola 95 a Lucillio - *«refert quare et quemadmodum fiant»*. Le stesse azioni possono essere turpi od oneste; dipende dal fine e dal modo in cui vengono realizzate. Appunto, il fine con cui ci si rivolge a uno strumento, e il modo con cui lo si utilizza, costituiscono ambiti che interpellano l'uomo, e tutto ciò che interpella invoca anche una

appare obiettivabile al pensiero umano. Cfr. F. Cavalla, *L'origine e il diritto*, FrancoAngeli, Milano 2017, pp. 18-22.

[37] Come è stato osservato, «è la consapevolezza del limite costitutivo dell'uomo – questo sapere il proprio non sapere – ciò che rende il soggetto responsabile nel senso più radicale del termine. Ciò che è dato all'uomo in questa consapevolezza del proprio limite è infatti la possibilità di accettarsi come tale nella propria costitutiva finitezza, di accogliere il proprio trovarsi ad esistere, di acconsentire al proprio essere nato» (L. Illetterati, *Figure del limite. Esperienze e forme della finitezza*, Verifiche, Trento 1996, p. 95).

[38] F. Zanuso, *Introduzione*, in F. Zanuso (a cura di), *Diritto e desiderio*, FrancoAngeli, Milano 2015, p. 22. «L'uomo viene all'esistenza, non per sua volontà, né per suo progetto. Ma nell'esistenza deve volere, deve progettare. Deve farlo riflettendo problematicamente. Cosa significa esistere alla luce della radice filosofica del termine? Stagliarsi, eccettuarsi, differenziarsi» (*Ibid*em).

responsabilità[39]. Siamo dunque, ancora – o forse più che mani – nella sfera della vita etica della persona umana, la quale è chiamata, vista la pervasività, la potenza e la novità delle innovazioni di cui stiamo parlando, a una vigilanza rafforzata: anche su di sé.

Nelle pagine che seguiranno esaminerò brevemente alcune sfide salienti sottese a due ambiti particolarmente significativi per indagare scenari e criticità dello sviluppo tecnologico-informatico contemporaneo: il Metaverso, da un lato; l'Intelligenza Artificiale, dall'altro. Di qui mi concentrerò su tre modalità attraverso le quali l'uso delle tecnologie digitali può accentuare la vulnerabilità umana, veicolando forme più o meno mascherate di violenza. Si tratta di tre diverse forme di "discriminazione" sulle quali a mio avviso è opportuno mantenere desta l'attenzione – tanto dei progettisti, quanto dei venditori, quanto soprattutto dei fruitori del mondo digitale.

In chiusa, vorrei tornare a ribadire l'importanza di un modello etico "pratico" che a mio avviso, pur nella sua semplicità, può aiutare a favorire un mutamento prospettico utile a mantenere un'attenzione sulla tecnologia intesa come strumento per l'essere umano e non sull'essere umano: la *restorative ethics*.

4. Potenzialità, miraggi e labirinti - I: brevi cenni su alcune sfide legate all'Intelligenza Artificiale

Il ricorso sempre più ampio a tecnologie informatiche avanzate (come l'Intelligenza Artificiale) o telematiche (si pensi alla connettività permanente in cui le nostre vite sono immerse, tramite varie *devices* tecnologiche) è certamente foriero di ampie potenzialità. Nel contempo, come detto, non sfuggono anche possibili risvolti di "compressione" dell'umano, che preconizzano o già realizzano esiti alienanti, strettamente connessi alla presenza di strumenti capaci di elevata pervasività sulle vite delle persone "immerse" nelle tecnologie informatiche. Tutto ciò si pone a danno di elementi fondamentali, come la *privacy*, e, ancor più, della stessa libertà, essendo tali tecnologie non di rado in grado di aumentare il livello di controllo sulle vite dei loro fruitori[40].

Non mancano, poi, scenari che evidenziano il rischio che un ampio (ed eccessivamente fiducioso) ricorso a tecnologie di IA faccia da amplificatore di errori anche

[39] Il tema è stato ripreso anche da papa Leone XIV: si veda, ad esempio, il messaggio ai partecipanti alla seconda *Conferenza Annuale su intelligenza artificiale, etica e governance d'impresa* (Roma, 19-20 giugno 2025). Osserva il Pontefice: «Insieme al suo straordinario potenziale di recare beneficio alla famiglia umana, il rapido sviluppo dell'intelligenza artificiale solleva anche questioni più profonde riguardanti l'uso corretto di tale tecnologia nel generare una società globale più autenticamente giusta e umana. In tal senso, pur essendo indubbiamente un prodotto eccezionale del genio umano, l'intelligenza artificiale è "innanzitutto uno strumento" (Papa Francesco, *Discorso alla Sessione del G7 sull'Intelligenza Artificiale*, 14 giugno 2024). Per definizione, gli strumenti rimandano all'intelligenza umana che li ha prodotti e traggono molta della loro forza etica dalle intenzioni delle persone che li impugnano. In alcuni casi l'intelligenza artificiale è stata utilizzata in modi positivi e perfino nobili per promuovere una maggiore uguaglianza, ma esiste anche la possibilità che venga usata male per un guadagno egoistico a spese altrui o, peggio ancora, per fomentare conflitti e aggressioni». (Si veda, per il testo integrale: https://www.vatican.va/content/leo-xiv/it/messages/pont-messages/2025/documents/20250617-messaggio-ia.html).

[40] Come osserva Mason Marks, «When firms acquire a dominant share of biopower, influencing enough traits in sufficiently large populations, they achieve biosupremacy, which this Article defines as monopolistic power over human behavior. Biosupremacy is a Digital Age analog of monopoly power. While monopoly power gives firms the ability to raise prices and exclude competitors within specific markets, biosupremacy enables firms to exert control, by shifting social norms over large swaths of human behavior, yielding influence that cuts across markets and entire industries» (M. Marks, *Biosupremacy: Big Data, Antitrust, and Monopolistic Power Over Human Behavior*, in «55 U.C. Davis Law Review», 513 (2021), pp. 513-589.

gravi, con conseguenze particolarmente pesanti: si parla, in questo caso, di un mondo informativo e informatico "inquinato" da un insieme di notizie e dati non reali, eppure verosimilmente presentati come tali, per effetto di una "contaminazione" di informazioni frutto di elaborazioni dell'IA. A mano a mano che i modelli di IA apprendono sempre più dati "contaminati" da *output* precedenti di IA, la qualità e l'affidabilità dei modelli futuri potrebbero degradarsi, fino al rischio di un *model collapse*[41]. Il paradosso è che i dati antecedenti al 2022 verrebbero considerati "puliti", mentre quelli successivi risulterebbero inquinati da una serie di alterazioni amplificate, con un effetto simile a quello del noto gioco del "telefono senza fili", eppure esponenzialmente più ampio e potenzialmente devastante.

Non solo: in un contesto di "bombardamento informatico" e di crescente "passività" e carenza critica nei confronti di ciò che viene passato dal *web*, è evidente che vi sono molti pericoli insiti in un sistema di trasmissione di informazioni in cui è sempre più difficile distinguere ciò che esiste nella realtà da ciò che è una mera rappresentazione virtuale e digitale di un mondo verosimile, e però inesistente, o, peggio, veicolante informazioni false. Oltre alle "allucinazioni" dell'IA – termine con cui viene designato un fenomeno in cui un modello linguistico di grandi dimensioni (LLM), spesso un *chatbot* di AI generativa, percepisce modelli o oggetti inesistenti o impercettibili agli osservatori umani, creando *output* privi di senso o del tutto imprecisi – c'è il serio rischio che l'essere umano sia continuamente colpito a sua volta dal rischio di "allucinazioni". Ci troveremmo quindi di fronte a un'umanità continuamente esposta a "miraggi digitali", fatto che colpisce non già la *manifestazione* del suo pensiero, bensì *inficia alla radice la formazione dello* stesso: ciò tocca, evidentemente, aspetti personalissimi e fondamentali della persona umana, intesa anche come soggetto politico e soggetto di diritti.

Il tema, dunque, non può sfuggire all'attenzione dei giuristi – e non necessariamente dei tecno-giuristi recentemente invocati come necessario sviluppo dei precedenti[42] – perché incide in modo estremamente rilevante sia sui diritti soggettivi, sia sugli strumenti di tutela degli stessi offerti dal diritto nella prassi[43].

Non è un caso, quindi, che su un tema dopo una serie articolata di lavori preparatori, l'Unione Europea abbia sentito l'importanza di legiferare in materia di tecnodiritto: come noto, nel marzo 2024 il Parlamento Europeo ha approvato il regolamento denominato *Artificial Intelligent Act*, atto normativo di carattere generale, «obbligatorio in tutti i suoi

---

[41] Si veda, per una prima infarinatura, https://www.theregister.com/2025/06/15/ai_model_collapse_pollution/ e, per una disamina più approfondita, J. Burden, M. Chiodo, H. Grosse Ruse-Khan, L. Markschies, D. Müller, S. O' hÉigeartaigh, R. Podszun, H. Zech, *Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training* (December 02, 2024), University of Cambridge Faculty of Law Research Paper No. 35/2024, Available at SSRN: https://ssrn.com/abstract=5045155 or http://dx.doi.org/10.2139/ssrn.5045155.

[42] Ci riferiamo all'idea dell'avvocato-ibrido recentemente proposta in P. Moro, *Tecnoetica forense. L'etica professionale dell'avvocato ibrido*, in F. Reggio, *Honeste Vivere. Percorsi filosofici per l'etica pubblica*, cit., pp. 211-222.

[43] Si vedano, a titolo esemplificativo, le riflessioni proposte, con un'attenzione ai risvolti etici, in P. Moro (a cura di), *Etica, Diritto, Tecnologia*, FrancoAngeli, Milano 2023. Cfr., altresì, sui mutamenti occorsi nella teoria e nella prassi giuridica, le emblematiche considerazioni proposte in G. Contissa, G. Sartor, *How the Law Has Become Computable*, in *Effective Protection of the Rights of the Accused in the EU Directives: A Computable Approach to Criminal Procedure Law*, Brill, Leiden 2022, pp. 27-41. Sull'evoluzione dei sistemi di risoluzione delle controversie per effetto dello sviluppo delle O.D.R. (*online dispute resolution*), soprattutto di seconda generazione, evidenziandone tanto le potenzialità quando i possibili esiti alienanti, si veda, in particolare, L. Mingardo, *Giustizia digitale alternativa: Scenari e riflessioni critiche sulle Online Dispute Resolution*, Primiceri, Padova 2020.

elementi e direttamente applicabile in ciascuno degli Stati membri (in base all'art. 298 del Trattato sul Funzionamento dell'Unione Europea)»[44]. È interessante osservare come al capitolo II, art. 5, il Regolamento inserisca una serie di proibizioni, relative ad applicazioni dell'IA di cui sono intraviste applicazioni particolarmente rischiose, tra cui, applicazioni dell'*Artificial Intelligence* vòlte a: (a) utilizzare tecniche subliminali, manipolative o ingannevoli per distorcere il comportamento e compromettere il processo decisionale informato, causando un danno significativo; (b) permettere lo sfruttamento delle vulnerabilità legate all'età, alla disabilità o alle condizioni socio-economiche per distorcere il comportamento, causando un danno significativo; (c) implementare sistemi di categorizzazione biometrica che deducono attributi sensibili (razza, opinioni politiche, appartenenza a sindacati, credenze religiose o filosofiche, vita sessuale o orientamento sessuale), a eccezione dell'etichettatura o del filtraggio di insiemi di dati biometrici acquisiti legalmente o quando le forze dell'ordine categorizzano i dati biometrici; (d) attuare *social scoring*, ossia la valutazione o la classificazione di individui o gruppi in base al comportamento sociale o a tratti personali, con conseguente trattamento svantaggioso o sfavorevole di tali persone; (e) valutare il rischio che un individuo commetta reati basandosi esclusivamente sul profilo o sui tratti della personalità, tranne quando viene utilizzato per aumentare le valutazioni umane basate su fatti oggettivi e verificabili direttamente collegati all'attività criminale.

L'elenco non è esaustivo, ma è significativo notare ciò che si intravede in controluce rispetto alle limitazioni previste dal Legislatore Europeo: i possibili risvolti di inaccettabile compressione della libertà e della dignità umane per effetto di un uso incontrollato delle tecnologie informatiche legate all'IA.

Se il sopra citato *Artificial Intelligence Act* è intervenuto con una regolamentazione piuttosto ampia, ciò è specchio di quanto questa tecnologia sia ormai diffusa: come è stato osservato, infatti, «l'Intelligenza Artificiale oggi viene utilizzata per automatizzare compiti e risolvere problemi complessi, trovando applicazione in un'ampia varietà di contesti: dalla ricerca scientifica al mercato azionario, dalla robotica alla giustizia, passando per l'industria dei giocattoli. L'IA è diventata sempre più rilevante nella nostra vita quotidiana e sta cambiando il mondo così come lo conosciamo»[45].

Un simile cambiamento, invero, non riguarda solo la potenzialità di avvalersi di strumenti di calcolo avanzato capaci di *supportare* attività e decisioni umane, nei più svariati ambiti, compreso quello giuridico[46]. Nello sviluppo contemporaneo dell'IA, infatti, un

---

[44] P. Moro, *Persona elettronica. Una finzione giuridica per l'intelligenza artificiale*, in «L'Ircorcervo», 1/2024, pp. 1-18, qui p. 2.

[45] C. Benetazzo, *Intelligenza artificiale e diritto: la sfida etica ed antropologica*, in «Journal of Ethics and Legal Technologies» – vol. 6(2) - December 2024, pp. 65-108, qui p. 66. L'impatto, poi, sull'ambito della conoscenza, è altrettanto significativo, come ben evidenziato in M. Palmirani, *Big data e conoscenza*, in «Rivista di Filosofia del Diritto», 1/2020, pp. 73 e ss.

[46] Cfr., per una prima lettura, G. Sartor, *L'intelligenza artificiale e il diritto*, Giappichelli, Torino, 2022; L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, *AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, in «Minds & Machines», 28 (2018), n. 4, pp. 689-707. Si vedano, altresì, i numeri monografici 1/2024 della rivista «L'Ircocervo» *(Intelligenza artificiale e scienze neuro-cognitive nel diritto: dalla simbiosi alla sostituzione/Artificial Intelligence and Neuro-cognitive Sciences in Law: From Symbiosis to Substitution)*, e il n. 2/2024 di «Journal of Ethics and Legal Technologies» *(Towards Responsible AI: Challenges and Perspectives Across Borders and Sectors*). Si distinguono, come è noto, varie tipologie di IA: (a) quella di tipo "predittivo", per la quale una macchina è in grado di fare previsioni sulla base di una combinazione di input precedenti

fronte particolarmente interessante, e nel contempo foriero di innumerevoli quesiti e di più di qualche preoccupazione, riguarda l'utilizzo della medesima secondo modalità pensate *in sostituzione* dell'attività umana, e non solo di quella di tipo operazionale o computazionale[47]. È il caso, questo, dell'IA *generativa*, la quale, avvalendosi di strumenti di *Machine Learning* anche particolarmente complessi, può arrivare a generare prodotti precedentemente inesistenti, tra cui, per esempio, testi, immagini, musica. Ampiamente mutuata, anche se non del tutto sostituita, in questo caso, è la creatività umana, e quindi una facoltà che fino a poco tempo addietro veniva considerata una peculiarità distintiva dell'umano[48].

Secondo uno studio a firma di Koivisto e Grassini, realizzato per confrontare la creatività umana con l'IA generativa, ai partecipanti era stato chiesto di generare usi insoliti e creativi per gli oggetti di uso quotidiano: ebbene, in media i *chatbot* di IA avrebbero superato i partecipanti umani. «Mentre le risposte umane includevano idee di scarsa qualità, i *chatbot* hanno generalmente prodotto risposte più creative»[49]. Tuttavia – precisano gli studiosi – «le migliori idee umane hanno comunque eguagliato o superato quelle dei *chatbot*. Se da un lato questo studio evidenzia il potenziale dell'IA come strumento per migliorare la creatività, dall'altro sottolinea la natura unica e complessa della creatività umana, che potrebbe essere difficile da replicare o superare completamente con la tecnologia IA»[50].

Se, dunque, almeno allo stato attuale delle conoscenze, occorre cautela nel ritenere che la capacità creativa umana sia sostituita (è riferito ad Albert Einstein l'aforisma secondo il quale «un giorno le macchine riusciranno a risolvere tutti i problemi, ma mai nessuna di esse potrà porne uno»), è altrettanto chiaro che gli scenari contemporanei mostrano

---

e di un'analisi delle tendenze e degli scenari attuali. Questo tipo di IA è già ampiamente utilizzato nel mondo aziendale; (b) quella "basata sulle anomalie", per la quale la macchina è programmata per riconoscere le regolarità e, pertanto, rilevare qualsiasi eccezione si presenti; (c) quella "basata sulle decisioni", che coadiuva, appunto, ad assumere decisioni in modo simile a quanto potrebbe fare un essere umano, ad esempio, attraverso classificazioni basate su caratteristiche. L'IA (d) generativa, invece, è studiata per realizzare prodotti innovativi, basandosi su una combinazione di dati acquisita ed elaborata ulteriormente. Si può utilizzare, per esempio, per produrre nuove immagini, testi o pattern musicali.

[47] Osserva Paolo Moro: «Già una decina d'anni orsono si avvertiva la possibilità che, attraverso l'utilizzo più efficace dell'intelligenza artificiale, i nuovi saperi del terzo millennio, come l'informatica, la robotica e le neuroscienze, avrebbero condotto alla costruzione di sistemi meccatronici complessi, che appaiono quasi come persone artificiali, in quanto imitano differenti facoltà umane: non soltanto ragionamenti, come la comprensione del linguaggio naturale, l'apprendimento automatico o la rappresentazione della conoscenza, ma anche comportamenti, come la capacità di decidere o la reazione a stimoli esterni, e in taluni casi anche emozioni, attraverso le espressioni del volto sintetico o altre azioni comunicative» (P. Moro, *Persona elettronica Una finzione giuridica per l'intelligenza artificiale,* in «L'Ircocervo», 23, 1/2024, pp. 372-389, qui, p. 373). Eppure, come osserva l'Autore, «i problemi etici e giuridici posti dall'intelligenza artificiale, talora erroneamente vista come un apparato univoco o un dispositivo indipendente, riguardano principalmente l'interazione umana con i sistemi robotici evoluti, i quali appaiono sempre più capaci di emulazione autonoma delle facoltà proprie del soggetto umano, del quale influenzano il modo di vivere ma dal cui controllo restano sempre più emancipati» (*Ibid.*).

[48] Cfr. K. Millet, F. Buehler, G. Du, M.D. Kokkoris, *Defending humankind: Anthropocentric bias in the appreciation of AI,* in «Computers and Human Behaviour», 143, 2023, 107707. https://doi.org/10.1016/j.chb.2023.107707.

[49] Cfr. M. Koivisto, S. Grassini, *Best humans still outperform artificial intelligence in a creative divergent thinking task,* in «Scientific Reports», vol. 13, 2023, Article number: 13601, *passim*.

[50] *Ibidem*.

potenzialità dell'IA tali da essere capaci di sostituire una vastissima serie di attività e processi per i quali era precedentemente necessaria una ampia compartecipazione umana[51].

Si aggiunge, così, un ulteriore capitolo del complesso rapporto uomo-macchina, che adombra sempre la possibilità che l'essere umano stesso possa essere messo da parte, surclassato, e finanche soggiogato da una sua creazione, alla stregua di un golem privo di controllo[52]. Come osserva dunque, puntualmente, Cristiana Benetazzo, «la sfida più pregnante posta dall'IA è antropologica, perché l'uso di strumenti di intelligenza artificiale rinvia necessariamente al rapporto tra l'uomo e la tecnologia, evocando la possibilità di sostituire il primo con una macchina, che è così umanizzata da essere persino costruita con le sembianze umane o, in ogni caso, con programmi che la fanno muovere e decidere come se si trattasse di un essere vivente»[53].

Non è obbiettivo del presente scritto aprire ampie parentesi sul tema dell'Intelligenza Artificiale, oggi già molto dibattuto: ciò che piuttosto preme qui è lasciar intravedere che esso è uno dei "fronti aperti" nei quali maggiormente si gioca una "sottile linea di confine" tra potenzialità e rischio, mostrando così, con singolare attualità, come l'ambivalenza del rapporto tra mondo umano e tecnica sia tale dal presentare un crinale rischioso, in cui lo strumento tecnologico, pur inizialmente concepito per dal "servire" l'umano, può facilmente finire per asservirlo[54].

5. Potenzialità, miraggi e labirinti - II: brevi cenni sul Metaverso

Queste considerazioni diventano ancora più pressanti se si leggono in combinazione con le questioni relative allo sviluppo del cosiddetto Metaverso, il quale, sebbene ultimamente sembri attrarre meno attenzione mediatica rispetto all'IA, non appare meno rilevante nel valutare scenari relativi all'impatto delle tecnologie digitali sul presente e sul prossimo futuro. È difficile invocare una definizione uniforme e condivisa di questo concetto: secondo uno studio realizzato in seno al Consiglio dell'Unione Europea, esso «è stato descritto come un costante e *immersivo* mondo virtuale tridimensionale nel quale le persone interagiscono attraverso un *avatar* per fruire di intrattenimento, compiere acquisti e condurre transazioni con cripto-valute, o condurre attività lavorativa da remoto (letteralmente "senza lasciare la loro sedia")»[55].

---

[51] Cfr., per alcune riflessioni critiche, dal taglio interdisciplinare: G. Piaia, R. Prete, L. Stefanutti (a cura di), *Intelligenza artificiale. Sviluppi futuri e tutela della persona (Philosophy, 2*), Triveneto Theology Press, Padova 2024.

[52] Com'è noto, la figura mitologica del golem è riconducibile alla cultura ebraica, ma anche, più in generale, al folklore medievale. Secondo la leggenda, chi viene a conoscenza della cabala, e in particolare dei poteri legati ai nomi di Dio, può fabbricare un golem, un gigante di argilla forte e ubbidiente, che può essere usato come servo, impiegato per svolgere lavori pesanti e come difensore del popolo ebraico dai suoi persecutori. Può essere evocato pronunciando una combinazione di lettere alfabetiche. In una delle leggende associate al golem, tuttavia, si narra anche del pericolo legato alla perdita di controllo del gigante. Cfr. M. Idel, *Il Golem – L'antropoide artificiale nelle tradizioni magiche e mistiche dell'ebraismo*, Einaudi, Torino 2006.

[53] C. Benetazzo, *Intelligenza artificiale e diritto: la sfida etica ed antropologica*, cit., pp. 68-69.

[54] Del resto, non mancano le implicazioni direttamente politiche legate all'uso dell'IA, a partire, ma non solo, dal suo pervasivo influsso sugli odierni sistemi di comunicazione. Cfr., a tal riguardo, R. Piroddi, *La comunicazione politica dell'era delle nuove tecnologie e dell'AI*, Eurilink University Press, Roma 2024.

[55] Council of the EU, analysis and research team, *Metaverse – virtual world, real challenges*, March 2022, p. 1. Più diffusamente, esso è stato anche così descritto, in uno studio curato da Christensen e Robinson per Analysisgroup: «While there is no agreed upon definition of the metaverse, one way to think about it is as an expansive network of digital spaces, including immersive 3D experiences in augmented, virtual, and

Volendo brevemente enucleare alcune caratteristiche essenziali del Metaverso potremmo evidenziare, i seguenti aspetti: esso propone (1) una *connettività immersiva* sul piano cognitivo e sensoriale, (2) *pervasiva* sul piano delle possibili attività che ivi si possono compiere (ludiche, comunicativo-relazionali, economico-finanziarie, giuridiche…) oltre che della frequenza con cui vi si potrà fare accesso per compiere tali attività, (3) altamente *virtualizzante* per la possibilità di proiettarsi su un mondo "alternativo", digitale e accessibile attraverso *avatars*, e, nel contempo, orientato ad (4) *apparire altamente realistico*, verisimile, a chi vi accede e interagisce come utente.

Alla luce di quanto osservato, non deve stupire che una sempre più vasta letteratura – scientifica, divulgativa e istituzionale – ne stia analizzando le caratteristiche, le potenzialità e, ovviamente, i rischi[56]. A titolo meramente esemplificativo, leggiamo quanto evidenziato nel già citato studio realizzato in seno al Consiglio dell'Unione Europea: «Il metaverso porta sia opportunità che rischi, la piena estensione e declinazione dei quali non è ancora chiara. Alcune problematiche relative a determinate *policies*, e alle loro possibili implicazioni, sono state identificate in diverse aree, tra cui la concorrenza, la protezione dei dati, le responsabilità giuridiche, le transazioni finanziarie, la *cybersecurity*, la salute, l'accessività e l'inclusività»[57].

Altri studi, maggiormente focalizzati sui profili etici legati alle tecnologie informatiche, evidenziano uno spettro rischi ancora maggiori, che si giocano su una pluralità di piani, distinti ma interconnessi, come quelli istituzionale, commerciale e interpersonale, ponendo sfide che toccano non solo i diritti fondamentali, bensì anche la salute psico-fisica degli utenti, oltre che la qualità delle relazioni e interazioni che, su una molteplicità di livelli, si possono "giocare" nel metaverso[58].

6. Tre forme di attacco alla dignità umana attraverso il digitale

Non si può negare che siano potenzialmente moltissimi gli ambiti nei quali gli sviluppi contemporanei del mondo digitale invocano una riflessione etica. Come preannunciato in

mixed reality, that are interconnected and interoperable so you can easily move between them, and in which you can create and explore with other people who are not in the same physical space as you. Some have referred to the metaverse as an "embodied internet" in which individuals will feel as if they are actually "present" in experiences and not simply looking at experiences through their screens. This means that interacting with the Internet (and the devices that provide access to the Internet) has the potential to be much more natural, incorporating modes of communication that include gesture and voice, such that individuals are not limited to typing or tapping. In addition, the metaverse is envisioned to be able to host almost all the activities we currently take part in (e.g., socializing, work, learning, entertainment, shopping, content creation, etc.) and make new types of activities possible as well)», L. Christensen, A. Robinson (eds.), *The potential global economic impact of the metaverse*, https://www.analysisgroup.com/globalassets/insights/publishing/2022-the-potential-global-economicimpact-of-the-metaverse.pdf (2022), accesso effettuato il 2 Sep 2022, p. 1.

[56] Essi sono stati recentemente ricordati, nel contesto delle attività culturali promosse a Bressanone dall'Università di Padova ai cosiddetti "corsi estivi", durante la *lectio magistralis* tenuta il 20 luglio 2023 dal prof. Pasquale Stanzione, Garante per la Protezione dei dati personali, intitolata *Habeas mentem: intelligenza artificiale, bioetica e tutela della persona*. Sempre il Garante ha promosso, il 30 gennaio 2023, un ampio convegno di studi in materia, intitolato *Il Metaverso tra utopie e distopie: orizzonti e sfide della protezione dei dati*, nel quale più diffusamente si è dibattuto delle sfide suscitate dagli scenari legati al Metaverso.

[57] Council of the EU, analysis and research team, *Metaverse - virtual world, real challenges*, March 2022, p. 11.

[58] Cfr., uno su tutti, R. Benjamins, Y. Rubio Viñuela, C. Alonso, *Social and ethical challenges of the metaverse Opening the debate,* in «AI and Ethics», 3, 2023, pp. 689-697.

precedenza, vorrei focalizzarmi sul fatto che la tecnologia digitale viene spesso associata anche al rischio di essere un fattore capace di creare e allargare distanze e gap tra esseri umani[59]. Non mancano, invero, spunti di riflessione nel dibattito contemporaneo che associano al digitale il pericolo di aumentare le disuguaglianze[60].

La letteratura contemporanea sembra concentrare la propria attenzione soprattutto su due versanti di questo fenomeno, spesso colti spesso sotto l'ombrello comune di *digital divide*: (a) la *digital discrimination*, che è la perpetuazione o manifestazione di *biases* (anche della vita reale) attraverso il digitale[61]; (b) il *digital divide* in senso proprio, ossia inteso come *digital exclusion* – che rappresenta una forma di esclusione sociale, o di mancata/inadeguata inclusività, per coloro i quali hanno difficoltà di accesso e fruizione dei servizi digitali (per età, istruzione, ceto), in forza della quale vengono a perpetuarsi e allargarsi vulnerabilità sociali[62].

Nel primo caso, ampiamente presente nel dibattito contemporaneo, soprattutto con riferimento a temi come le discriminazioni di genere o a matrice razziale[63], ci troviamo di fronte a una situazione in cui la tecnologia digitale è solamente un mezzo per la realizzazione di forme di discriminazione che possono nella realtà presentarsi anche a prescindere dal digitale stesso. Nel secondo caso, invece, è la stessa tecnologia digitale a costituire un *divider*, ossia una fonte di divisione/discriminazione nell'accesso a beni, servizi, conoscenze o informazioni tra soggetti dotati di oggettiva o soggettiva difficoltà ad accedere alla tecnologia stessa. Si pensi, per esempio, agli anziani che, privi di adeguata alfabetizzazione e strumentazione di accesso al digitale, si trovano sempre più in difficoltà, in molte società contemporanee, a compiere operazioni un tempo facilmente accessibili di persona o al telefono, e che oggi invece richiedono attività da svolgere tramite internet, talora con livelli di complessità ancora elevati per l'attuale cittadino medio.

A mio avviso vi è, tuttavia, una terza forma di discriminazione che può essere veicolata tramite la tecnologia digitale, e che ritengo vada tenuta opportunamente distinta dalle prime due pocanzi citate, ancorché possa con esse intersecarsi. Si tratta della (c) *digital de-responsabilization*.

È facile imbattersi in questo fenomeno nel contesto dell'utilizzo di piattaforme digitali e informatizzate: qui, volontariamente o meno da parte dei *provider* e dei gestori di

---

[59] Emblematico sul punto il Ted Talk di Sherry Turkle: https://www.ted.com/talks/sherry_turkle_connected_but_alone. Si veda anche, per un approfondimento, S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books, New York 2011.

[60] Cfr. L. Royakkers, J. Timmer, L. Kool *et al.*, *Societal and ethical issues of digitization*, in «Ethics and Information Technology», 20, 2018, pp. 127-142, https://doi.org/10.1007/s10676-018-9452-x; A. Grybauskas, A. Stefanini, M. Ghobakhloo, *Social sustainability in the age of digitalization: A systematic literature Review on the social implications of industry 4.0*, in «Technology in Society», 70, 2022, https://doi.org/10.1016/j.techsoc.2022.101997.

[61] Si veda, sul punto, N. Criado, J.M. Such, *Digital Discrimination*, in K. Yeung, M. Lodge (eds), *Algorithmic Regulation*, online edition, Oxford Academic, Oxford 2019), https://doi.org/10.1093/oso/9780198838494.003.0004, accessed 23 July 2023.

[62] Cfr., sul punto, D. Owen, *Digital divide*, in G. Mazzoleni, K. Barnhurst, K. Ikeda, H. Wessler, R. Maia (eds.), *The international encyclopedia of political communication*, 2016, doi:10.1002/9781118541555.wbiepc176; F.P. Jamil Marques, M. Coulart Massuchin, I. Mitozo, *Digital Divide*, in *The International Encyclopedia of Gender, Media, and Communication*, Wiley-Blackwell Publishing, Hoboken (NJ) 2020, pp. 1-7.

[63] Si veda, una su tutti, J. Daniels, *Rethinking cyberfeminism(s): Race, gender, and embodiment*, in «Women's Studies Quarterly», vol. 37, n. 1-2, 2009, pp. 101-124. doi:10.1353/wsq.0.0158.

tali strutture, invece che coadiuvare l'accesso a prestazioni, o servizi (anche di *customer care*), lo strumento digitale frappone un insieme di ostacoli intermedi tra la persona che ha un problema da risolvere (es. un'informazione da conseguire, un servizio da utilizzare...) e la possibilità effettiva di fruirne. Si va dall'utilizzo di software di IA (es. *chat bot*) per rispondere a quesiti, a *call-centers* digitalizzati, a chat informative, a strumenti di *do-it-yourself* digitale, rispetto ai quali diviene sempre più difficile non solo trovare un altro essere umano dall'altro capo della propria "connessione" ma anche trovare un autentico ascolto e un'autentica assistenza rispetto a un problema, che non sempre rientra nelle categorie e nei "bivi concettuali" precostituiti da chi ha realizzato il sistema digitale in questione. Le esperienze di queste tipologie di (dis)servizi sono molteplici, e facilmente accessibili a chiunque abbia fatto esperienza, per esempio, di voli cancellati, di esigenze di *customer care* legate ad acquisti o transazioni su internet, di call-center digitalizzati per la risoluzione di problemi legati a utenze, per esempio telefoniche.

Appare legittimo chiedersi se simili situazioni siano semplicemente frutto di una mera "autoreferenzialità digitale", irriflessa, specchio di strumenti pensati intorno al programmatore o al fornitore dei servizi, e non al fruitore. Non di rado, tuttavia, può sorgere il sospetto di trovarsi al cospetto di una serie di labirinti digitali artatamente predisposti, strutturati al fine di facilitare la *non-responsabilità* (interpersonale, sociale ma anche giuridica) di un determinato soggetto (per restare ai nostri esempi, di un fornitore di servizi).

Al di là delle diversità che caratterizzano le tre categorie di *digital discrimination* sinora enucleate, ciò che le accomuna è il fare della digitalizzazione e della informatizzazione un terreno atto non a rimuovere ma a perpetuare o allargare discriminazioni, con potenziali danni soprattutto per categorie dotate di maggiori vulnerabilità personali e sociali[64].

Ciò che invece le distingue è che se l'ambito della (a) *digital discrimination* semplicemente traspone sul piano digitale *dividers* che già esistono sul piano 'reale', la categoria della (b) *digital exclusion* fa del digitale stesso una linea entro cui si dipana una possibile discriminazione intorno all'accessibilità e fruibilità del mondo digitale e dei servizi che sempre più spesso sono esclusivamente (ed "escludentemente") affidati a tale tecnologia. Con crescente impatto sul piano etico, sociale e giuridico, la (c) *digital de-responsabilization* fa del digitale stesso uno strumento per *creare* barriere e percorsi deresponsabilizzanti[65].

Come ho avuto modo di esprimere anche in altre sedi, credo sia prioritario oggi ragionare sulla pensabilità di uno *human-based design*[66], anche se oggi è meglio parlare di *human-centered design,* a rimarcare che questo concetto non esclude in alcun modo le tecnologie basate sulla IA. Mi riferisco, con questa terminologia a una progettazione di servizi tecnologici disegnati intorno alle persone che devono accedervi e fruirne, con

---

[64] Cfr. M. Pérez-Escolar, F. Canet, *Research on vulnerable people and digital inclusion: toward a consolidated taxonomical framework*, in «Universal Access in the Information Society», 22, 2023, pp. 1059-1072, https://doi.org/10.1007/s10209-022-00867-x.

[65] Non a caso vengono invocati, in risposta, approcci e strumenti atti a potenziare la partecipazione personale e sociale alla vita pubblica, tanto nello spazio civico quanto nello spazio digitale, come antidoto alla divisione prodotta dalle tecnologie digitali. Cfr., sul punto, A. Volterrani, M. Storice, M.C. Antonucci, *Social and digital vulnerabilities: The role of participatory processes in the reconfiguration of urban and digital space*, in «Frontiers in Political Science», 4/2022, pp. 1-12, doi:10.3389/fpos.2022.970958.

[66] Cfr. F. Reggio, *Tecnologia per l'umano. Spunti di restorative ethics per una tecnologia digitale 'human centred'*, in Id., (a cura di), *Honeste Vivere. Percorsi filosofici per l'etica pubblica*, cit., pp. 133-172.

attenzione primaria alle loro esigenze e cura verso ciò che possa anzitutto evitare di esporli nelle loro possibili vulnerabilità, se non, ove possibile, cercare di porvi rimedio.

Questo concetto non va inteso come semplice teoria, nel senso di astrazione concettuale, bensì di una "visione" che fornisca ben precise "lenti" e "messe a fuoco" per la progettazione e l'utilizzo di strumenti digitali, avendo come cornice generale l'idea di una tecnica "*per* l'uomo" e non "*sull*'uomo".

7. Possibili spunti a partire da una prospettiva di restorative ethics

Tra le prospettive che, nella recente evoluzione culturale, anche giuridica, invocano e propongono un diverso sguardo e metodologie mirate, finalizzate anzitutto a tutelare l'essere umano nella sua *suitas* e nella sua dimensione relazionale, una delle più interessanti è a mio avviso la *restorative justice*. Nata intorno a una riflessione radicalmente critica della giustizia penale moderna e contemporanea, rispetto alla quale propone un ripensamento complessivo[67], questa prospettiva è a mio avviso particolarmente interessante per uno "sguardo d'insieme" che la caratterizza, e che travalica il settore penalistico. Incentrata sulle dimensioni della partecipazione, della riparazione e del consenso[68], la prospettiva *restorative* appare particolarmente feconda per intavolare una riflessione sulle attitudini che possano aiutare a prevenire il rischio che un essere umano si trovi schiacciato, strumentalizzato, o ridotto a *res*, nel contesto delle sue interazioni sociali. In secondo luogo, nella sua attitudine "pratica", essa invita concretamente a intervenire in modo costruttivo e riparativo qualora una simile reificazione si fosse verificata.

In questo senso, la prospettiva *restorative* si inserisce in un interessante filone di importanti sollecitazioni che, anche nel dibattito giuridico, hanno evidenziato il rischio di una "disumanizzazione" e l'esigenza di prospettive e metodologie disegnate intorno all'importanza di tutelare e valorizzare l'essere umano come soggettività relazionale. Riportare al centro la persona nella sua dimensione intersoggettiva ivi compresa la sua capacità di essere parte attiva dei processi decisionali che la riguardano, è un tema oramai trasversale a diversi ambiti di riflessione teorica e sperimentazione pratica, a partire dalla macro-categoria della *participatory justice*[69]. Tra gli ambiti che maggiormente hanno posto l'enfasi sulle dimensioni dell'autonomia, della partecipazione, del dialogo e del consenso tra *stakeholders*, quali forme atte a maggiormente proteggere e valorizzare la soggettività

[67] Cfr., sul punto, H. Zehr, *Changing Lenses. A new Focus on Crime and Justice*, Herald Press, Scottsdale 1990; M. Wright, *Justice for victims and offender*s, Open University Press, Philadelphia 1991; W. Cragg, *The practice of punishment. Towards a theory of restorative justice*, Routeledge, London-New York 1992.

[68] «La restorative justice – o giustizia rigenerativa – è un approccio alla giustizia che considera il reato principalmente in termini di lesione alle persone, da cui scaturisce in capo all'autore l'obbligo di porre rimedio alle conseguenze dannose della sua condotta. A tal fine, la prospettiva restorative mira a realizzare un coinvolgimento attivo di vittima, offensore, del loro rispettivo entourage e della stessa comunità civile nella ricerca di soluzioni – possibilmente concordate – atte a far fronte all'insieme di bisogni scaturiti a seguito del reato» (F. Reggio*, La nave di Milinda. La Restorative Justice fra conquiste e sfide ancora aperte,* in C. Sarra, F. Reggio, *Diritto, metodologia giuridica e composizione del conflitto,* Primiceri, Padova 2020, pp. 11-100, qui p. 21).

[69] Si vedano, ad esempio, alcuni interessanti e pionieristici studi: G. Stephens, *Participatory justice: The politics of the future*, in «Justice Quarterly», March 1986, vol. 3, no. 1, pp. 67-82 (16); R. Axelrod, *The evolution of cooperation*, Basic Books, New York 1984. Interessante, invece, a livello istituzionale, quanto elaborato in Law Commission of Canada, *Towards Participatory Justice: A Focus on People and Relationships*, Ottawa, Canada 2000. Cfr., altresì, S. Liebenberg, *Participatory Justice in Social Rights Adjudication*, in «Human Rights Law Review», vol. 18, n. 4, 2018, pp. 623-649.

relazionale delle persone[70], non possiamo dimenticare il ruolo giocato, negli ultimi anni, dalla riflessione prospettica e metodologica sugli strumenti autonomi di soluzione della controversia, in particolare la negoziazione[71] e la mediazione[72].

Eppure, tra questi diversi modelli, è proprio la *restorative justice* a mostrare, a mio avviso, la più ampia capacità di influenzare sempre più vasti ambiti di indagine e di azione[73], tanto che sta sempre più evolvendosi in un vero *social movement*[74]. Ancora nel 2007, Gerry Johnstone, intervenendo sul tema, aveva rilevato che la proposta *restorative* si è estesa al di fuori dell'ambito della giustizia penale, secondo un moto "discendente" che l'ha portata a offrire prospettive e metodi per i conflitti nelle scuole, nei luoghi di lavoro e nelle condotte potenzialmente lesive che appartengono alla vita quotidiana e non assumono rilevanza penale, e, parallelamente, secondo un movimento "ascendente", che ha portato a considerare le potenzialità offerte dal paradigma *restorative* in ambiti più estesi, come la violenza politica, ampie violazioni dei diritti umani[75], e ingiustizie storiche o sociali di vasta scala[76].

---

[70] Cfr., per considerazioni filosofico-giuridiche, in dialogo con la tradizione della "prospettiva processuale del diritto", F. Reggio, *Il diritto tra 'conversione del conflitto in controversia' e 'abilitazione al dialogo'. La prospettiva processuale del diritto alla prova degli strumenti ADR consensuali*, in S. Fuselli, P. Moro, E. Pariotti (a cura di), *Universa Universis Patavina Libertas. Filosofia del Diritto a Padova tra innovazione e tradizione. Per celebrare gli 800 anni dell'Ateneo*, Primiceri, Padova 2022, pp. 59-104.

[71] Cfr., per una prima rassegna, l'ormai classico, R. Fisher, W. Ury, B. Patton, *Getting to Yes: negotiating Agreement Without Giving In*, Penguin Books, New York 1991, e, più recentemente, R.D. Benjamin, *The Natural History of Negotiation and Mediation: The Evolution of Negotiative Behaviors, Rituals, and Approaches*, 2012; G. Amira, *The World of Negotiation*, World Scientific Publishing, New York 2016; G. Richard Shell, *Bargaining for Advantage: Negotiation Strategies for Reasonable People*, Penguin Books, New York 1999. Nella dottrina italiana, con un approccio critico verso la concezione "tecnorazionalista" della prima elaborazione harvardiana, si veda M. Antonazzi, *Il negoziato psicologico*, Eurilink University Press, Roma 2017; M. Antonazzi, *La teoria della negoziazione cognitiva*, in F. Reggio, C. Sarra (a cura di), *Diritto, Metodologia Giuridica e Composizione del Conflitto*, Primiceri, Padova, 2020, pp. 181-216.

[72] Sulla mediazione come uno dei possibili volti della giustizia partecipativa, cfr. M.A. Foddai, *Participatory Justice and Mediation Toward a New Model of Justice*, in «Soft Power», 4, 2016, pp. 127-143; sulla mediazione nella sua dimensione relazionale e nel suo volto di riumanizzazione della giustizia, non si può mancare di ricordare il contributo della mediazione umanistica, per cui si rinvia agli studi di Jacqueline Morineau, recentemente scomparsa (J. Morineau, *Esprit de La Mediation*, Eres, Toulouse 1998). La mediazione come spazio di riattivazione di risorse relazionali, nelle dimensioni di *empowerment* e *recognition*, è stata oggetto di elaborazione in R. Baruch Busch, J. Folger, *The promise of Mediation: Responding to Conflict through Empowerment and Recognition*, Jossey Bass, San Francisco 1994. Sulla mediazione come spazio vòlto ad attivare una via dialogica per la conversione del conflitto in controversia, nel quale riveste importanza centrale il ruolo dell'argomentazione giuridica, si rinvia a F. Reggio, *Concordare la norma. Gli strumenti consensuali di soluzione della controversia in ambito civile: una prospettiva filosofico-metodologica*, Cleup, Padova 2017. Sul punto, vanno ricordati anche i pionieristici studi di John William Cooley, per cui si veda, quale compendio, J.W. Cooley, *The Mediator's Handbook. Advanced Practice Guide for Civil Litigation*, National Institute for Trial Advocacy Press, Boulder (CO) 2006. Cfr., altresì, in una prospettiva influenzata dalla pragma-dialettica, S. Greco, *Dal conflitto al dialogo. Un approccio comunicativo alla mediazione*, Maggioli, Firenze 2021.

[73] G. Johnstone, D. Van Ness, (2007). *The meaning of restorative justice*, in G. Johnstone, D. Van Ness (eds.), *Handbook of restorative justice*, Willan Publishing, London 2007, pp. 5-23.

[74] Lo evidenziano, recentissimamente, Fernanda Fonseca Rosenblatt (che ringrazio sentitamente per la condivisione) e Craig Adamson, in F. Fonseca Rosenblatt, C. Adamson, *Non-encounter restorative justice interventions – now what?,* in «Contemporary Justice Review», 1, 2023, 1-18, doi: 10.1080/10282580.2023.221671.

[75] G. Johnstone, *Restorative justice: Ideas, Values, Debates* (2nd ed.), Routledge, London 2011, p. 144.

[76] Sul punto, si rinvia, a titolo esemplificativo, a D.C. Emling, *Institutional racism and restorative justice: Oppression*

L'estensione del riferimento alla prospettiva *restorative* travalica però lo stretto settore della "giustizia": anzi, è a mio avviso opportuno evitare di diluire eccessivamente la *restorative justice* associandola a un insieme troppo variegato ed eterogeneo di ambiti di intervento (es. ambiente, cambiamento climatico…), con il rischio che essa diventi un riferimento endossale, vago, e, ancor peggio, "di moda".

Diviene qui particolarmente interessante la proposta di estrapolare dal paradigma *justice* una *restorative ethics*: già da tempo, invero, uno dei "padri" di questa prospettiva, Howard Zehr, aveva suggerito l'idea che la prospettiva restorative potesse divenire una "visione che guidi e sostenga le nostre vite", valicando il confine della riflessione penalistica[77]. Ponendosi quasi a mo' di novella etica delle virtù, la "visione" invocata da Zehr fa della prospettiva *restorative* un'attitudine verso il mondo: un "modo di guardare" a se stessi e alle reti di relazioni in cui si è immersi così da porre debita enfasi su tre elementi: relazionalità, responsabilità, rispetto. Termini forse vaghi, e sicuramente bisognosi di ulteriore problematizzazione, ma non per questo privi di pregnanza nel contesto della vita etica contemporanea.

Ciò che la prospettiva *restorative* insegna, per esempio, è ad "allenare lo sguardo" a cogliere, onorare l'umano, ponendosi in condizione di attivare dinamiche di rispetto, riconoscimento, relazione, visibili già nelle scelte comunicative. Cercare di prevenire i rischi di riduzione dell'umano a oggetto, attivandosi per rimediare qualora ciò sia accaduto, non è un mero slogan, se insegna a ponderare tanto i motivi ispiratori quanto il possibile impatto delle proprie azioni, mettendo in guardia dal rischio di solipsistiche e irresponsabili autoreferenzialità, così presenti nella 'vita praticata' contemporanea.

I nuclei prospettici intorno a cui l'idea di *restorative ethics* si addensa, infatti, aiutano a mantenere focalizzata l'attenzione su aspetti fondamentali che rischiano di finire compressi e dimenticati, vieppiù, come si è visto, nel mondo tecno-centrico della nostra contemporaneità, perché invitano a un'attenzione e a una sollecitudine rafforzate, come appunto ben indicato dalla metafora delle "lenti" e della loro "messa a fuoco". Per esempio, l'attenzione che la prospettiva *restorative* pone ai bisogni delle persone porta con sé necessariamente che la *ratio* di un certo prodotto o servizio non sia disgiunta dalle esigenze dei potenziali fruitori, a partire da quelli più vulnerabili.

Per restare quindi al mondo digitale, il *design* di quel determinato prodotto o servizio deve essere pensato in modo *user friendly*, e, soprattutto, finalizzato a permettere all'utente di conseguire ciò che verosimilmente orienta la sua azione. Tutto questo invita a ponderare il linguaggio, la grafica, le opzioni sul tavolo in modo da non fare del sistema digitale un labirinto che pone ostacoli tra l'esigenza della persona e le possibilità del suo soddisfacimento. Ciò vale soprattutto per le persone che hanno maggiori difficoltà, proprio per evitare forme di *digital discrimination*. Parallelamente, l'ottica *restorative* impone, come si è detto, di valutare l'impatto che azioni e comunicazioni sortiscono sulle persone, in particolare se vulnerabili, e quindi insegna a non limitarsi a un punto di vista puramente operativo o funzionale nel *design* di un prodotto o un servizio. Essa richiede, quindi, una sorveglianza, per la quale occorre, per esempio, figurarsi le possibili conseguenze legate

and privilege in America, Routeledge, London 2020; F. Reggio, *Two wrongs don't make one right. Memory, History and Rebalancing Actions. A reading on 'cancel culture' through the lens of a Restorative Approach*, in «Calumet – intercultural law and humanities review», 16, 2023, pp. 1-28.

[77] Cfr. H. Zehr, *Restorative Justice Beyond Crime. A Vision to Guide and Sustain our Lives*, in «Verifiche», 2/2019, pp. 11-12, ora presente, tradotto in italiano, in F. Reggio (a cura di), *Honeste Vivere*, cit., cap. II.

all'utilizzo di un determinato strumento e non solo e non tanto in termini di responsabilità strettamente legale (sulla quale molto spesso si può agire in via preventiva, e, in un certo senso deresponsabilizzante, attraverso una serie di clausole legate a *disclaimers* o a manifestazioni del consenso previe rispetto all'accesso a un determinato servizio), bensì in termini anche di responsabilità etica e sociale (e perché no, anche ambientale) in senso più ampio.

L'attenzione alla persona che un modello di *restorative ethics* richiede si rende visibile anche nella riflessione sui processi e sulle modalità operative: rispetto a questi ultimi appare fondamentale aver cura della dimensione partecipativa e dialogica sottese al rispetto della persona nella sua dignità e relazionalità. Per restare agli esempi citati con riferimento alle forme di *digital deresponsabilization*, un servizio digitalizzato che disincentivi la possibilità di raffrontarsi con un altro essere umano intorno a un bisogno o a una peculiare esigenza, costituisce un meccanismo che impedisce "di metterci la faccia" e che, dietro il paravento di una automazione che agevoli la personalizzazione della risposta, in realtà crea barriere alla possibilità di vedere tutelata una propria prerogativa o un proprio diritto.

Per citare un esempio concreto, un servizio di assistenza clienti che offra assistenza stradale in caso di guasto o panne, ma che renda difficile, all'utente medio, attivare la possibilità di una chiamata di soccorso, costringendolo anzi a compilare questionari online digitando dal cellulare, e per di più in una situazione di emergenza, costituisce un sistema alienante e che non assolve al compito di rispondere ai bisogni di una persona in un momento di difficoltà. L'assenza di una comunicazione diretta, poi, aumenta il senso di solitudine e di alienazione di chi, appunto, si trova ad avere necessità di un'assistenza pronta, rafforzando il senso di pericolo e di difficoltà. In questo caso, il designer del servizio in questione, nonché i gestori, mancano completamente di offrire ciò che è nella *ratio* stessa dell'assistenza, e per di più si esimono dalla più basilare forma di responsabilità, ossia, rispondere a qualcuno, prima ancora che rispondere di qualcosa. Il fatto poi che il sistema digitalizzato utilizzi strumenti di automazione che standardizzano e non personalizzano la risposta, priva di quella attenzione al caso concreto nella sua contestualità che costituisce invece un vero e proprio *leitmotiv* dell'attitudine *restorative*.

Gli esempi si possono moltiplicare, e sono certo che il lettore che voglia richiamarsi alla sua personale esperienza ne troverà diversi e validi (magari anche solo nella difficoltà di vedersi consegnata una corrispondenza o un prodotto acquistato con e-commerce). Penso che tutti convergeranno nella direzione di evidenziare il bisogno di una maggiore attenzione e cura per le persone e le relazioni umane. Forse questo può apparire un'istanza irenistica o buonista, però in concreto non credo si possa negare che la carenza di una simile attenzione e cura sia alla base di molti dei malesseri, per non dire degli scenari disumanizzanti, in cui si dibatte la nostra inquieta contemporaneità.

8. Ciononostante, homo sum

Ciò che preme, al fine delle nostre riflessioni, e l'attualità ricorsiva della considerazione di Seneca che ho già richiamato in precedenza: "le stesse cose possono essere turpi od oneste" – *per* la persona o *sulla* persona: "dipende dal come, e dal fine per cui siano state fatte" (o progettate). L'ottica *restorative* insegna quella sollecitudine verso la persona concreta, i suoi bisogni, le sue difficoltà ed esigenze, che difficilmente un'automazione può avere, ma la cui

carenza, allo stato attuale, spesso si verifica per una mancanza etica, se non per una mancanza di intelligenza umana, di chi l'ha posta in essere.

Se si concorderà con queste considerazioni, probabilmente si converrà anche con l'idea che Metaverso da un lato e IA dall'altro – senza contare le evidenti interazioni tra questi ambiti – non sono dunque altro che esempi contemporanei di un problema antico e contemporaneo, che non si risolve *nella*, *con la* ed *entro la* tecnologia, bensì invoca una riflessione critica sul mondo umano.

Qui però la questione potrebbe tingersi di sfumature oscure: forse aveva ragione Jacqueline Morineau quando sosteneva che da tempo l'Occidente vive una (*macro*)crisi, intesa come «un periodo di perdita dei valori e dei punti di riferimento in cui risulta sempre più difficile dare senso alla vita»[78]? All'interno di tale cornice, le tante (*micro*)crisi che si susseguono non sarebbero altro che conferme di una situazione di sostanziale smarrimento nella quale l'uomo contemporaneo è immerso: momenti nei quali tale fenomeno, da carsico, riaffiora, in modo più o meno drammatico e potente.

Eppure, "crisi" non è un termine che va colto esclusivamente nella difettività (di senso, di riferimenti, di sicurezze) che esso invoca, nello stato di "indebolimento" che essa provoca rispetto ad uno *status quo* precedente. Nella cultura classica tale concetto è ambivalente, e questo è legato allo stesso verbo *krinein* e al suo duplice significato: «da un lato, la locuzione allude all'idea di frantumazione, sgretolamento e deriva dall'indoeuropeo *sqerei* (da cui i termini "scorie", "avanzi"), presente nell'espressione ellenica *skorìen poiein*, che vuol dire "mandare in rovina"; dall'altro lato, il verbo indica specificamente il "dire giustizia" e significa propriamente "giudicare distinguendo", ossia unificare ciò che opponendosi si distingue»[79].

In questo unificare c'è anche il "ripristinare", il "rigenerare" ciò che merita di essere preservato. Ciò che quindi va *restored* – perché la tecnologia digitale sia *human-centered* e non *dehumanizing* – è anzitutto la capacità dell'essere umano stesso di recuperare un'autentica sollecitudine verso il suo simile (e non solo): *homo sum, humani nihil a me alienum puto* per citare l'imperituro verso di Terenzio. A ricordarci che spesso la prima disumanizzazione non è operata dalla tecnologia, bensì da chi fruisce della stessa con uno sguardo dimentico dell'umanità propria e altrui.

Forse la vera sfida, dunque, non è umanizzare le macchine, ma mantenere umano l'uomo stesso. Nonostante la tecnologia. Nonostante il digitale. Nonostante l'uomo.

[78] J. Morineau, *Esprit de La Mediation*, cit., p. 63.
[79] P. Moro, *Alle origini del Nómos nella Grecia classica. Una prospettiva della legge per il presente*, FrancoAngeli, Milano 2014, p. 20.

Algorithmic processing and AI bias; using overfitting to reveal rather than perpetuate existing bias[a]

Lydia Farina[*] and Anna-Maria Piskopani[†]

Abstract

In questo articolo analizziamo l'*overfitting* dell'IA nell'elaborazione algoritmica per mostrare come esso sia correlato a casi di iniquità o distorsione dell'IA e come si combini con fenomeni sociali complessi, quali gli effetti di looping, per mantenere ed esacerbare le distorsioni esistenti. Discutiamo le normative esistenti e proposte in materia di IA che tentano di affrontare questo pregiudizio per cogliere le tendenze e le priorità dominanti. Infine, suggeriamo che, sebbene l'attenzione della letteratura attualmente si concentri sulle conseguenze negative dell'*overfitting*, esso può essere utilizzato come strumento diagnostico per individuare le disuguaglianze sociali sottostanti e, in quanto tale, portare a usi alternativi dell'analisi dell'IA per smascherare l'ingiustizia sociale piuttosto che esacerbarla. Questo articolo fornisce un ulteriore supporto teorico alle recenti opinioni presenti nella letteratura che suggeriscono che l'elaborazione algoritmica può essere utilizzata per diagnosticare e monitorare i pregiudizi; evidenziando l'interazione con gli effetti di looping, fornisce anche un'ulteriore motivazione per utilizzare l'*overfitting* come primo passo verso la mitigazione dei pregiudizi storici.

Keywords: overfitting, regolamentazione IA, valutazione diritti umani, equità algoritmica, discriminazione algoritmica.

In this paper we analyse AI overfitting in algorithmic processing to show how it relates to cases of unfairness or AI bias and how it combines with complex social phenomena such as looping effects to maintain and exacerbate existing bias. We discuss existing and proposed AI regulation attempting to address this bias to pick up dominant trends and priorities. Finally, we suggest that, although the focus of the literature currently falls on the negative consequences of overfitting, it can be used as a diagnostic tool for detecting underlying social inequalities and, as such, lead to alternative uses of AI analytics to expose social injustice rather than exacerbate it. This paper provides further theoretical support to recent views in the literature suggesting that algorithmic processing can be used to diagnose and monitor bias; by highlighting the interaction with looping effects, it also provides additional motivation to use overfitting as a first step towards mitigation of historical prejudice.

[a] Received on 22/05/2025 and published on 09/12/2025.
[*] University of Nottingham, e-mail: lydia.farina@nottingham.ac.uk.
[†] University of Nottingham, e-mail: anna-maria.piskopani@nottingham.ac.uk.

Keywords: overfitting, AI regulation, human rights assessments, algorithmic fairness, algorithmic discrimination.

Introduction

AI overfitting is described in at least two different ways: 1) as an observed tendency in AI based algorithms to take shortcuts to achieve a given task by ignoring some data which do not agree with common trends[1] or 2) as a tendency to inaccurately fit patterns based on limited data to more generalised cases, hence 'over fitting'[2]. To give a more refined definition, overfitting happens when algorithmic models make predictions based on regularities discovered in the training data which do not match the world from which the data is taken[3].

Overfitting can exacerbate bias in certain domains where algorithmic processing is used to generate outcome predictions after training such as in the practices of predictive policing or predictive healthcare. In predictive policing algorithmic processing generates predictions such as a percentage of crime rate detection in a particular area[4]. The algorithmic prediction is used as a justification for sending additional officers to that area. In turn, this results in more crimes being detected in that area as the larger number of officers correlates with more crimes being detected. The updated detection crime rate of the area is then fed back as data to the algorithm, to be used for subsequent predictions. In certain cases of predictive healthcare, the algorithm predicts healthcare needs by calculating incurred healthcare costs[5]. Areas where healthcare costs are higher, are then allocated more funding on the assumption that costs incurred track healthcare needs of the population living in these areas. Using these specific contexts as case examples we show that allowing algorithmic predictions to dictate policy in these contexts not only maintains bias but exacerbates it.

Even though the risks and harms associated with algorithmic processing in specific contexts e.g. in predictive policing have been identified for several years, such practices are still being used[6]. In more recent years, human rights organisations such as Conseil of Europe, the European Agency for Fundamental Rights, international organisations such as United Nations published reports and resolutions identifying ways in which AI based

[1] J. Shane, *You Look Like A Thing and I Love You*, Headline Publishing Group, Wilfire 2019.
[2] See K.P. Burnham, D.R. Anderson, *Model selection and multimodel inference*. 2nd ed., Springer-Verlag 2002; J.S. Russell, P. Norvig, *Artificial Intelligence; A Modern Approach*, 3rd ed.., Pearson Education Limited 2016.
[3] D.L Poole., A.K. Mackworth, *Artificial Intelligence; Foundations of Computational Agents*, 3rd ed., Cambridge University Press, Cambridge UK 2023, p. 297.
[4] See K. Hao, *Police Across the US are training Crime Predicting Ais on Falsified Data*, «MIT Technology Review», 13 February 2019, Police across the US are training crime-predicting AIs on falsified data | MIT Technology Review; A. Larson, J. Angwin, Bias in *Criminal Risk Scores is Mathematically Inevitable, Researchers Say*, «ProPublica», 30 December 2016, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.
[5] See S. Gupta, *Bias in a Common Health Care Algorithm Disproportionately Hurts Black Patients*, «SCI. NEWS», Oct. 24 2019. https://www.sciencenews.org/article/bias-common-health-care-algorithm- hurts-black-patients.
[6] See M. Degeling, B. Berendt, *What is wrong about Robocops as consultants? A technology-centric critique of predictive policing*, «AI & Society», n. 33, 2018, pp. 347–356, https://doi.org/10.1007/s00146-017-0730-7.

decisions violate human rights (identified as non-discrimination and data protection rights). The UK published an AI regulation White Paper[7] where it describes its approach to AI challenges as well as a roadmap to effective AI assurance ecosystems. The UK ICO (Information Commissioner's Office) published guidance for public services aiming to evaluate whether such violations or risks to equality rights take place along with a roadmap to effective AI assurance ecosystems[8].

In what follows we first show how overfitting relates to cases of unfairness and to phenomena discussed in social ontology such as 'looping effects' to maintain and exacerbate existing bias especially in cases where discrimination is based on characteristics which are not protected by law[9]. We then discuss existing and proposed AI regulation attempting to address bias arising from algorithmic processing to pick up dominant trends and priorities. Finally, we suggest that, although the focus of the literature currently falls on the negative consequences of overfitting, it can be used as a tool to reveal underlying social inequalities and social injustice when AI is used for social welfare, policing etc and, these findings can be repurposed and further explored by other scientists such as cognitive scientists to understand in depth implicit biases and stereotypes and lead to further interdisciplinary research[10]. This paper provides further theoretical support to recent views in the literature suggesting that algorithmic processing can be used to reveal, diagnose and even monitor bias as a first step towards mitigation of historical prejudice[11].

1. *Overfitting, feedback loops and looping effects*

Algorithms are mathematical constructs tasked with picking up patterns available from their training data. In the cases discussed in this paper, AI bias is a direct consequence of the tendency shown by algorithms to ignore some of the data (relating to minority characteristics) and privilege others, that is the ones associated with the dominant pattern, to reveal the dominant pattern found in the data (overfitting via shortcut). So, the first type of AI overfitting relates to excluding data which do not reflect dominant trends[12]. For example, if we task an algorithm with selecting amongst CVs to fill in a job vacancy, the

[7] Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, 2023, CP 815, retrieved 10 April 2024 from: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[8] Information Commissioner's Office, *Guidance in AI and Data Protection*, updated 2023. Retrieved on 10 April 2024, available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/.

[9] For a discussion of looping effects, see I. Hacking, The looping effects of human kinds, in D. Sperber and A. J. Premack. (eds.), *Causal cognition; A Multidisciplinary Debate*. Clarendon Press, New York 1995, pp. 351-394.

[10] For a recent example of using algorithmic processing to provide evidence for bias against the poor in social networks see G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings*, «AI & Society», n. 39, 2024, pp. 617–632. https://doi.org/10.1007/s00146-022-01494-z.

[11] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, «AI & Society», n. 38, 2023, pp. 747–76.

[12] J. Shane, *You Look Like A Thing and I Love You*, cit.

data it is trained on will have an impact on the decision it reaches.[13] Typically, the data training set includes examples of successful CVs for similar positions. If the majority of the successful CVs share a characteristic e.g. sex this can be considered as an essential characteristic by the algorithm so that successful CVs not including this characteristic are ignored. This leads to a generation of a model which does not represent all training data. Instead, the model is based on the data which give the faster route to the preset solution interpreted as a cluster of necessary characteristics found in most successful CVs. In practice this means that even if the training data set includes minority characteristics, through algorithmic processing, this data is being ignored. This leads to inaccurate models of real environments[14].

In addition, a second type of overfitting is associated with known errors such as using unrepresentative sampling and it can be combined with the former type of overfitting to exacerbate AI bias. When the algorithm is trained on limited data, its outcome predictions can only be accurate when applied to limited cases such as the ones reflected by the limited data. However, if the prediction generated by the algorithm, is used as a generic model to be applied to different environments - so environments not reflected in its training data, the prediction is inaccurate as it is fitted or applied to environments not reflected in its training set (overfitting via overextending). For example, we may train an algorithm tasked to recognise facial expressions by using images of facial expressions from 30 students of a particular age group or of a particular social background studying at a specific university. If we then use our algorithm to predict the facial expression of an individual who is not represented in the training data, we are asking our algorithm to overextend. AI overfitting occurs with all types of learners such as decision trees or neural nets and it partially depends on the number of training examples; the more training examples are included in the data, the lesser degree of this type of AI overfitting occurs[15]. Both types of overfitting discussed above lead to biased predictions or the creation of inaccurate causal models and as such can be considered as instances of AI bias.

Importantly, the quality of the training data has an impact on the level of overfitting. For example, quality data could lead to the generation of more accurate models in the sense of the models matching real environments or contexts. The accuracy would depend on whether the models accurately reflect causal patterns in real environments. To ensure that algorithmic processing would generate accurate models, these causal patterns must be reflected in the data we feed to the algorithm. However, this presupposes that:

1) we already have a good idea and understanding of these causal patterns (condition 1) and
2) we have a mechanism e.g., a test or screening, which checks that the data fed to the algorithm reflect these patterns and do not reflect racist, sexist etc. patterns (condition 2).

[13] Here we are using a hypothetical scenario but for a study on gender bias in hiring see A. Peng, B. Nushi, E. Kıcıman, K. Inkpen, S. Suri, E. Kamar, *What you see is what you get? The impact of representation criteria on human bias in hiring*, «Proceedings of the AAAI Conference on Human Computation and Crowdsourcing», vol. 7, n. 1, 2019, pp. 125-134.
[14] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.
[15] See J.S. Russell, P. Norvig, *Artificial Intelligence, A Modern Approach*, cit.

Both these requirements need to be satisfied and so are necessary but not sufficient for the training data to be considered as quality data. The second condition, that is the requirement that we check that the training data does not reflect discriminatory patterns, is a necessary condition for quality data because in its absence algorithmic processing can discover patterns in the data even when there are no causally relevant patterns to be discovered within. By relevant here we mean patterns showing a causal relation between features or properties included in the data and the outcome/task we set the algorithm to perform. For example, if the task is to predict whether the roll of a dice will come up as 5 or not, it is not relevant whether the thrower of the dice is wearing a batman suit or pyjamas when throwing the dice or whether the day is Tuesday, Wednesday or Thursday. However, if these data are fed into the training set of the algorithm, it is possible that a pattern can be discovered which predicts that the dice will come up as 5 if today is Tuesday and the thrower is wearing pyjamas. Generally speaking, in AI overfitting the algorithm will come up with a pattern which will be fitted 'over' the data even when this pattern may not reveal any causally relevant patterns.

As many people would expect the use of algorithmic processing to infuse objectivity and accuracy into this process, this discrimination is amplified; because of the 'veil of objectivity' associated with using AI systems rather than humans in these contexts, there are less checks and tests on the decisions/predictions generated compared to decisions/predictions generated by human agents[16].

Ultimately if our priority is to generate accurate models of the world, algorithmic processing must be primed towards accuracy. In simple contexts and environments this is easily achieved. On the other hand, when the model we are after relates to complex environments, algorithmic processing cannot guarantee accuracy. As such, we need to ensure that algorithmic processing is used in the right context and not presuppose that it is a tool that can be used to generate models in every context[17]. In addition, when algorithmic processing is used to make predictions or decisions on complex matters where many variables are causally relevant and some of these relevant variables are morally salient e.g. determining parole, predictive policing, prison sentences, mortgage applications etc., fairness must also be primed if one wants to argue that these tasks are appropriate tasks for algorithmic processing[18].

Bias maintained via overfitting can be further exacerbated by "looping effects" where an interactive loop is created between the data we use to train the algorithm and the outcome predictions generated by the algorithm. To show how looping effects can

[16] The mistaken assumption here is that the algorithm cannot have any biases relating to race, sex, class etc. because it is not an agent, even when these biases are endemic in the data it is trained on. For a discussion of AI bias and its implications on human users, see M. Glickman, T. Sharot, 15 November 2022, *Biased AI systems produce biased humans*. https://doi.org/10.31219/osf.io/c4e7r.

[17] L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, cit.

[18] See B. Giovanola, S. Tiribelli, *Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms*, «AI & Society», n. 38, 2023, pp. 549-563. https://doi.org/10.1007/s00146-022-01455-6; B. Green, Y. Chen, *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments*, «FAT* 2019 - Proc 2019 Conference on Fairness, Accountability, Transparency», 2019, pp. 90-99. https://doi.org/10.1145/3287560.3287563; P. Hacker, *Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law*, «Common Market Law Review», n. *55*, 2018, pp. 1143-1185. https://doi.org/10.54648/COLA2018095.

exacerbate the effects of overfitting, consider an example relating to healthcare costs. An algorithm is tasked with calculating a percentage of healthcare needs by focusing on healthcare costs; it predicts that white patients or patients living in specific areas - the ones where people spend more on healthcare costs, will need additional care. An immediate result may be that more resources are allocated towards healthcare needs of white patients or to patients living in these specific areas. A subsequent result is that people who consider themselves as members of these groups or classifications e.g., white patients or patients living in these specific areas, become aware that other members of their group spend more on healthcare costs. The possible consequences of this awareness are either that patients change their behaviour to conform with what other members of the group are doing, do nothing, or change their behaviour to stop conforming with other members of the group. In this paper we focus on the first possibility as this can have a detrimental effect in exacerbating existing socio-economic imbalances and further increasing the negative effects of AI overfitting. The patients living in this area now have a larger selection of healthcare facilities or activities in the area which gives them the opportunity to spend more funds on healthcare costs. This creates an interacting loop where the more funds are spent towards healthcare costs, via the subsequent algorithmic prediction, the more funding is allocated to that area for example towards healthcare facilities, so that patients end up spending even more in that area. However, as we hope the example shows, by using algorithmic processing to predict healthcare needs, the wrong conclusion is being reached in that an area with fewer healthcare needs is flagged up as an area with more healthcare needs[19].

One obvious reason why the above happens is that we are drawing the wrong conclusion from actual data because we are allowing wrongful associations between healthcare costs and healthcare needs. Instead, it is often argued that we should train the algorithm with data showing actual healthcare needs and levels of health. However, doing this may be complicated because some of this data will be protected by data protection law as special category of data for very good reasons or because we have insufficient data to determine healthcare needs.

In the third part of this paper agreeing with Zajko[20] we suggest that, algorithmic processing in general, and overfitting in particular, is a tool that could be used to alleviate rather than exacerbate socioeconomic bias in certain specific cases.

2. Regulatory frameworks addressing AI bias with focus on overfitting

As AI research and innovation is advancing rapidly, it is not clear that current legislation is sufficient or effective in protecting individuals from the negative consequences of algorithmic processing. The existing legal framework in the EU offers some opportunities to fight discriminatory effects of algorithmic processing through data protection law (especially Article 22 GDPR), consumer law, criminal and administrative law (regarding fair procedures) or the provisions of anti-discrimination law. At the same time, there are significant difficulties in applying these in practice. For example, in relation to equality laws,

[19] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, «Science», 25 October 2019, 366, 6464, pp. 447-453. 10.1126/science.aax2342.
[20] See M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, «AI & Society», 36, 2021, pp. 1047-1056. https://doi.org/10.1007/s00146-021-01153-9.

the prohibition of indirect discrimination does not provide a clear and easily applicable rule; it needs to be proven that a seemingly neutral policy or measure disproportionately affects a protected group and is therefore *prima facie* discriminatory. This means that statistical evidence is needed to prove that such a disproportionate result is taking place. Moreover, non-discrimination laws apply to defined groups sharing certain protected characteristics, but AI systems could negatively affect groups of people who do not share such protected characteristics for example people living in specific neighbourhoods or people who are associated with particular political parties[21].

Another limitation of the existing legal framework is that data protection law does not apply to algorithmic processing of non-personal data. For example, predictive models can include non-identifiable data that are outside the scope of data protection law. As mentioned above, sometimes processing personal data is necessary to identify AI bias and discriminatory affects, in the sense that we may need to use algorithmic processing on personal data to prove discrimination is taking place[22]. This latter consideration reveals a tension between data protection rights and the need to eradicate discrimination and bias. Most importantly, individuals must be aware of the use of algorithms when these are used in decisions or predictions that impact on their lives. However, most of these harms such as unfair discrimination, or data protection breaches are for a long time invisible to the people subjected to them[23].

As the existing regulatory framework is struggling to address the issues posed by the algorithmic processing, measures have been proposed from human rights organisations to protect among other public interests, fairness, and human rights. In relation to predicting policing, the European Agency for Fundamental Rights (FRA) highlights that the quality of training data and other sources associated with bias need to be mandatorily assessed by the users of these autonomous analytics systems[24]. It also suggests that the outputs of algorithmic systems become the basis for updated algorithms and that assessments are needed both before and during the use of algorithmic processing.

The above feed into the discussion about the necessity of AI regulation addressing AI systems' risks. The AI Act Regulation (EU 2024/1689) laying down harmonised rules on artificial intelligence entered into force on 1 August 2024, and will be fully applicable 2 years later on 2 August 2026, with some exceptions.[25] This regulatory framework has a lot in common with the data protection framework. It attempts to establish obligations for AI providers, deployers and users depending on the level of risk the AI systems can generate, and the adverse impact caused by AI systems on fundamental protected rights e.g., dignity, protection of personal data, right to non-discrimination, employment rights, rights of persons with disabilities, presumption of innocence. When such risks and impacts are identified, AI applications should be classified as high-risk.

---

[21] See F. Zuiderveen Borgesius, F., *Discrimination, artificial intelligence, and algorithmic decision-making*, Directorate General of Democracy, Council of Europe, 2018; U. Peters, *Algorithmic political bias in artificial intelligence systems*, «Philosophy & Technology», n. 35, 2022. https://doi.org/10.1007/s13347-022-00512-8.

[22] See M.J. Kushner, J.R. Loftus, *The Long Road to Fairer Algorithms: Build models that identify and mitigate the causes of discrimination*, «Nature», 578, 2020, pp. 34-38.

[23] See C. Véliz, *Privacy Is Power*. Melville House, London, 2021, p.39.

[24] See FRA. European Union Agency for Fundamental Rights, *Bias in algorithms. Artificial Intelligence and Discrimination*, Publications Office of the European Union, Luxembourg 2022.

[25] Available online at: https://artificialintelligenceact.eu/the-act/.

Examples of high-risk AI systems are the ones used for selection of persons in educational or training institutes, for recruitment purposes, promotion decisions, social score, credit score evaluations or creditworthiness of persons (e.g. profiling). In the public sector, examples of high-risk AI applications are the ones used for decisions relating to social benefit entitlement, predictive policing, crime analytics, emergency services, or evaluation of healthcare needs (Annex III). To mitigate the risks, AI systems should be used only if they comply with certain mandatory requirements, such as risk management, use of high-quality and relevant data sets, maintaining technical documentation, record-keeping transparency, the provision of information to deployers, human oversight, robustness, accuracy and cybersecurity as well as compliance with the EU legislation. Risk management systems should consist of a continuous, iterative process that is planned and run throughout the entire lifecycle of using AI systems in high-risk areas and it should be regularly reviewed and updated to ensure its continuing effectiveness.

One of the AI Act's main objectives is to mitigate discrimination and bias in the development, deployment, and use of "high-risk AI systems". The AI Act takes under consideration cases of overfitting similar to the ones analysed in this paper and has related provisions. According to art.10 of the AI Act, high-risk AI systems which make use of techniques involving the training of AI models with data, should be developed with quality criteria on the basis of training, validation and testing data sets. Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. These data sets should have the appropriate statistical properties, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used. Data sets shall consider, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioural, or functional setting within which the high-risk AI system is intended to be used. In addition, according to 10 (5), for the purposes of ensuring bias detection and correction, processing of sensitive data for the use of high-risk AI systems is exceptionally permitted, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons.

According to article 12 (5), high-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures. Outputs of AI systems could be influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination.

As additional safeguards, deployers of high-risk AI systems which are also bodies governed by public law (e.g. police) as well private operators that provide services such as banking and insurance and private entities providing public services linked to tasks in the public interest such as in the areas of education, healthcare, social services, housing, administration of justice have the obligation to the extent the deployer exercises control over the input data, they must ensure that input data is relevant and sufficiently representative in view of the intended purpose of the high-risk AI system (art. 26). They also carry out a fundamental rights impact assessment prior to using the AI system and determine measures to mitigate the identified risks arrangements through human oversight, complaint handling and redress procedures (art. 27). Deployers can also involve relevant

stakeholders e.g. representatives of people who are expected to be affected by these AI systems and co-design measures to mitigate the risks. Furthermore, the affected people have the right to lodge a complaint with the market surveillance authority and the right to request an explanation if the output of certain high-risk system produces legal effects or significantly affects and impacts their health, safety, or fundamental rights (art. 86).

As shown above, during recent years there have been several regulatory attempts to identify the problem caused by AI bias, identify obligations for developers and users, assess the life cycle of these systems and suggest best practices to minimise their socially negative impacts. Unlike the EU approach followed by countries as Canada and possibly Latin American countries, the UK published AI White Paper (2023) did not aim to suggest new AI specific regulation; instead, the UK elected to use a context specific, principles-based framework. It released guidelines to empower regulators, allowing for statutory action to be called upon when necessary. The White Paper was put in and the UK government published its response (February 2024). One of the key issues raised by the respondents is that the government's focus on innovation, does not allow room for sufficient focus on AI-related risks, such as bias and discrimination.

All existing and proposed regulations discussed above share the assumptions that algorithmic processing increases efficiency, augments existing capabilities, and has beneficial results for the economy. These regulations and policies also accept that algorithmic processing may replace some human decision-marking processing, maintain discrimination or be a threat to humans whilst the effectiveness of any suggested remedies remains questionable[26].

3. Reconceptualising overfitting and AI system deployment

As shown in the previous chapter, regulating the use of algorithmic processing is currently a work in progress. Broadly speaking, up to now overfitting has been perceived as a technical issue which can be fixed by a technical solution, similar to a machine malfunction. AI assurance scholars have attempted to create standardised methods to detect bias arising from algorithmic processing[27]. Examples of such solutions are bias screening software[28], preprocessing[29] or rigorous testing[30]. These suggestions try to minimise any discrimination observed in algorithmic predictions and decisions by including tests or screenings after the development of the algorithm and before it is released for use.

---

[26] See G. De Gregorio, S. Demkova, *The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe* (SSRN Scholarly Paper 4712096), «Social Science Research Network», 2024, https://doi.org/10.2139/ssrn.4712096; For more detail see the government response on https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response#fn:45.

[27] See F. Batarseh, J. Chandrasekaran, L. Freeman, *An introduction to AI assurance*, in A. Feras, F. Batarseh, L. Freeman (eds.), *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*, Elsevier Science & Technology, Amsterdam 2022.

[28] See C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, «Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)», Association for Computing Machinery, New York 2021, pp. 666-677, https://doi.org/10.1145/3442188.3445928.

[29] F. Kamiran, T. Calders, *Data preprocessing techniques for classification without discrimination*, «Knowledge and Information Systems», 33, 2012, pp. 1-33.

[30] See M.J. Kushner, J.R. Loftus, *The Long Road to Fairer Algorithms*, cit.

However, and as Johnson argues, «There are no purely algorithmic solutions to the problems that face algorithmic bias»[31]. Partly this is because AI overfitting does not only occur when we develop faulty mathematical constructs. It occurs regardless as it based on inferential reasoning; we use the algorithm as a mechanism extracting patterns from data we feed into it. When the algorithm picks up patterns in the data, these often are the result of structural discrimination or intersectional discrimination. When this algorithm is used to generate decisions in the form of prediction purposes, it reinforces these inequalities in societies. Whilst reviewing current and proposed legislation we mentioned that in some cases to prove discrimination is taking place one needs to provide statistical evidence that discrimination is taking place. Here we would like to suggest that algorithmic processing can be used to detect this discrimination and so play the part of statistical evidence proving discrimination. Risk and human rights assessments imposed by AI regulation can play a substantial role by providing the framework and justification for using algorithmic processing to detect rather than perpetuate bias. However, for this to happen these risk and human rights assessments should be developed in a way that incorporate interdisciplinary approaches and related academic research and serve the purposes of accountability and social control[32]. External (i.e., second party) tracking mechanisms and independent (i.e., third-party) oversight that blend the process and the outcome of the algorithm systems can surpass any bureaucratic uses[33]. We could use algorithmic processing to identify where bias is occurring within different contexts especially in cases where this harm is invisible and undetected for long periods[34].

In what follows, borrowing from recent research we give an example of how overfitting could be included as an important step in the exercise of detecting bias and revealing improper decision-making processes.

3.a Example of using overfitting to detect bias and faults in decision making process

A Guardian investigation[35] in 2019 found that UK local councils used algorithmic processing on data held on claimants of housing and council tax benefit to determine the likelihood these claims were fraudulent by using "risk-based verification". These systems were designed to perceive and reveal groups of people as types of risks[36]. However, as the investigation revealed, most of the cases deemed high risk by the software were in fact lower risk, and as a result, benefit claims were wrongly delayed. The algorithms drew on

[31] G.M. Johnson, *Algorithmic bias: on the implicit biases of social technology*, «Synthese», 198, 2021, p. 9943, https://doi.org/10.1007/s11229-020-02696-y.
[32] See I.J. Monteiro, *The Need for Responsible Use of AI by Public Administration: Algorithmic Impact Assessments (AIAs) as Instruments for Accountability and Social Control*, in J. Goossens, E. Keymolen, A. Stanojević (eds.), *Public Governance and Emerging Technologies*, Springer, Cham 2025. https://doi.org/10.1007/978-3-031-84748-6_9.
[33] See A. Brandusescu, R. Sieber, *Design versus reality: assessing the results and compliance of algorithmic impact assessments*, «Digital Society», vol. 4, art. 64, 2025, https://doi.org/10.1007/s44206-025-00221-7.
[34] See M. Zajko, *Conservative AI and social inequality*, cit.
[35] S. Marsh, *One in three councils using algorithms to make welfare decisions*, «The Guardian», 15 October 2019, https://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits.
[36] L. Dencik, A. Hintz, J. Redden, H. Warne, *Data Scores as Governance: Investigating uses of citizen scoring in public services*, Project Report, Data Justice Lab. Cardiff University, 2018.

and used data about people who make use of social services; the data itself was biased through the over-representation of a particular part of the population making use of social services often as a result of being marginalised and resource poor[37]. Overfitting in this instance not only revealed a problem with the data itself but also with the task we set the algorithm to perform and with how the algorithm was used in the decision-making process. It revealed that the decision-making process did not follow proper procedure.

To prevent such occurrences, we suggest following the framework and principles suggested by recent research conducted by the Centre for Data Ethics and Innovation and ICO. The Centre for Data Ethics and Innovation independent report titled *Review into bias in algorithmic decision-making*[38], analyses paradigms in four different areas as recruitment, financial services, policing and social services and highlights the importance of involving all stakeholders e.g. decision makers, industry, policy makers and society to determine whether the overall decision-making processes are biased. It suggests the following as important factors in alleviating bias from algorithmic decision making when they are used by organisations for decision making:

A. Understanding the capabilities and limits of those tools.
B. Considering carefully whether individuals will be fairly treated by the decision-making process that the tool forms part of.
C. Making a conscious decision on appropriate levels of human involvement in the decision-making process.
D. Putting structures in place to gather data and monitor outcomes for fairness.
E. Understanding their legal obligations and having carried out appropriate impact assessments

These recommendations formed the key principles of the AI Playbook for UK Government[39]. In addition, ICO in its *AI fairness considerations across the AI lifecycle* guidance (Annex A), proposes the following methodology when overfitting is detected to evaluate the data and its relation to the algorithm and the model generated:

1) examine the data's context (particularly whether there is an appropriate representation of groups in the training data);
2) evaluate the values that are assigned to the features before someone feeds them into the model,
3) tweak the model by tuning its hyperparameters; and
4) fit the most appropriate algorithm to the data. The algorithm might not be the appropriate one, so other algorithms must be tested as well as their performance.

[37] This was also observed in the use of the SyRI (system risk indication) system implemented by the Dutch government to identify potential fraudulent beneficiaries, see M. Bekkum, F. Borgesius, *Digital welfare fraud detection and the Dutch SyRI judgment*, «European Journal of Social Security», 23, 2021, pp. 323-340, https://doi.org/10.1177/13882627211031257.

[38] Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making*, 2020. Retrieved 10 April 2024, from: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making.

[39] Government Digital Service, *Artificial Intelligence Playbook for the UK Government*, February 2025. Retrieved 1 August 2025 from: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government.

The guidelines and framework included above should be used as soon as AI overfitting is discovered to re-evaluate the whole decision-making process. In other words, overfitting is an indicator that a proper procedure was not followed when determining the decision-making process. Although detecting overfitting is not a simple procedure, empirical evidence of overfitting can be obtained when the generalisation error of a model generated by the algorithm is higher than what is expected by the algorithm's performance in training[40]. Our suggestion matches one of the aims of the Centre for Data and Innovation review, namely that algorithms can enable the identification and mitigation of systematic bias in cases where doing so would be challenging for human agents. If the task of the algorithm changes from generating predictions matching dominant trends to revealing patterns of dominant trends and generating models based on these patterns this could help identify these patterns as discriminatory.

Applying the framework in the example discussed above we would need to: a) justify the use of data processing in that area e.g. evidence for fraud in the area b) conduct a community consultation to identify benefits and risks of using data processing c) reevaluating the hypothesis that the length of time of using the benefits constitutes an indication of fraud, d) consider the wider context by consulting with experts e.g. social workers e) consider consequences of using algorithmic processing in this case and identify any remedies for harms caused from the outcomes generated by the algorithm f) re-evaluate the wider context and address wider structural societal conditions which maintain the need for housing and council tax benefits.

The purpose of this paper is not to demonstrate a concrete example of using algorithmic processing to detect discriminatory patterns but to motivate decision makers and AI developers by changing the current narratives surrounding the use of AI systems. More specifically, technical experts in AI should be tasked with developing algorithmic tools which can detect hidden discrimination and provide concrete evidence for it. Agreeing with recent literature we accept that AI experts are to some extent, responsible for foreseeable consequences of use of AI technology[41]. Our suggestion can be used as one way AI experts can accept this responsibility by developing tools to detect and identify rather than exacerbate discrimination. This different perspective could contribute to a new narrative applying to AI systems and their social benefits which could potentially demystify and de-demonise them.

In addition, using AI overfitting could provide evidence to initiatives aiming to identify unfairness in key parts of individuals' lives such as in healthcare systems where discrimination is invisible even when training is based on partially filtered data. Using algorithmic processing to reveal patterns associated with healthcare contexts could reveal patterns of discrimination by using proxy features such as post codes, political affiliations etc., data which is not currently protected by law. The AI Act will be implemented in the coming years so that risk management and human rights impact assessments become obligatory in AI systems placed in EU countries. As the constant revaluation of these

[40] C. Aliferis, G.J. Simon, *Overfitting, Underfitting and General Model Overconfidence and Under Performance Pitfalls and Best Practices in Machine Learning and AI*, in G.J. Simon, C. Aliferis (eds.), *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, Springer, Cham 2024, pp.477-525.
European Parliament and the Council of European Union, *Artificial Intelligence Act (Final Draft)*, 2024. Accessed 1 April 2024, available at: https://artificialintelligenceact.eu/the-act/.

[41] M. Hedlund, E. Persson, *Expert responsibility in AI development*, «AI & Society», n. 39, pp. 453-464, https://doi.org/10.1007/s00146-022-01498-9.

systems will be obligatory, using algorithmic processing to reveal, explore and monitor existing bias can provide useful evidence for societies invested in ameliorating social and economic inequalities.

Our suggestion is motivated by recent initiatives addressing social or economic discrimination which aim to bring in the centre of this discourse the larger social, economic, and political ecosystem creating discriminatory practices and procedures. Discrimination maintained by algorithmic processing and exclusionary automated systems represent one element of this larger ecosystem. The identification of problems and any potential solutions should be linked with measures of reflexivity in relation to this ecosystem[42]. Our suggestion to use AI overfitting to reveal hidden discrimination could be a practical example of how this reflexivity of the ecosystem takes place in practical terms.

In a similar way it is suggested that we need to evaluate how data used in training map onto recent social, economic, cultural, and political changes on a global scale[43]. More recent views emphasize the need to evaluate the fairness of algorithmic practices by embedding them in social practices instead of focusing on evaluating outcome predictions by mathematical constructs[44]. These approaches pay particular attention to structural inequalities that are reproduced in the algorithmic decision-making process and suggest that these mathematical constructs must extend to relational or structural factors associated with the specific task. We interpret this to mean that the algorithms must be tasked with producing more accurate models of the training data. Agreeing with these recent holistic approaches, we suggest using algorithmic processing to reveal hidden discrimination. This will provide bottom-up evidence which can be helpful to several initiatives suggesting that we should seek to build alternative bottom-up infrastructures to empower marginalised groups and avoid such harms in the future[45].

Societies invested on positive social change can evaluate the use of AI systems during the process of impact assessment which is a process highly prioritised in the AI Act and becomes mandatory from August 2026. Decision makers in critical areas such as welfare and social care, healthcare, transportation, housing and planning, education, policing, and public safety could use any overfitting findings for further policy making and scientific use. An organisation possessing hard data on all these different domains, can provide insight into discriminatory practices endemic in the data tracking non-protected characteristics to provide hard evidence as the first step in the process of addressing them. The next step would be to come up with solutions addressing such inequalities by combining human and machine intelligence[46]. For example, in predictive policing instead

[42] S.P. Gangadharan, J. Niklas, *Decentering technology in discourse on discrimination*, «Information, Communication & Society», vol. 22, n. 7, 2019, pp. 882-899, https://doi.org/10.1080/1369118X.2019.1593484.

[43] L. Dencik, A. Hintz, J. Redden, E. Treré, *Exploring Data Justice: Conceptions, Applications and Directions*, «Information, Communication & Society», vol. 22, n. 7, 2019, pp. 873-881, https://doi.org/10.1080/1369118X.2019.1606268.

[44] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, «Philosophy & Technology», vol. 35, n. 90, 2022, https://doi.org/10.1007/s13347-022-00584-6; S. Holm, *The fairness in algorithmic fairness*, «Res Publica», 2022, https://doi.org/10.1007/s11158-022- 09546-3.

[45] S. Costanza-Chock, *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*, «Proceedings of the Design Research Society 2018», 3 June 2018, available at SSRN: https://ssrn.com/abstract=3189696.

[46] G. Mulgan, *Artificial intelligence and collective intelligence: the emergence of a new field*, «AI & Society», n. 33, 2018, pp. 631–632, https://doi.org/10.1007/s00146-018-0861-5.

of allowing confirmation bias to create feedback loops around specific neighbourhoods[47], algorithmic processing could be used as a diagnostic tool to reveal areas needing urgent development[48].

4. Conclusion

To conclude we hope the above discussion shows that even if AI overfitting is mathematically inevitable, its negative effects can be reframed as catalysts for change and opportunities for public awareness and scientific research. As the review of recent literature shows, we urgently need a constant and persistent effort to detect hidden discrimination prior to, during and after use of algorithmic processing. The fundamental impact assessments that are or will become obligatory (AI Act) for decision makers (deployers of the high risk systems) can assist to raise awareness both for the quality of data included in these training sets and of the adverse effects of AI applications trained on them. If these assessments are implemented in practice by serving purposes of transparency, accountability and social control, they could prevent harmful results to both individuals and communities. A further and more crucial step would be to act on revealed social problems and attempt their resolution.

In this paper we provide support to recent views suggesting that overfitting can be used as a diagnostic tool in the development of AI systems by signifying whether the requirements for responsible use of AI have been met[49] and to views suggesting that new narratives can lead to more responsible and transparent AI practices[50]. As recent attempts to regulate algorithmic processing show, in previous years harm was caused from maintaining and exacerbating structural and historical inequalities. In many cases this harm was invisible to the individuals and communities subjected to it. Recent legislation suggests steps to restrict or make its use safer in certain contexts. This is confirmation that algorithmic processing should not be used in all contexts for all tasks. We support these initiatives and accept that algorithmic processing can be used more radically as a diagnostic tool to detect and reveal hidden structural and historical bias and provide evidence for pre-existing prejudice[51]. By highlighting the interaction with looping effects, we provide additional motivation to use overfitting as a first step towards mitigation of historical prejudice. It remains a choice for decision makers whether they wish to combat or maintain

---

[47] A. Babuta, M. Oswald, *Data Analytics and Algorithms in Policing in England and Wales: Towards A New Policy Framework*, «RUSI Occasional Paper», 2020, available at: https://rusi.org/publication/occasional-papers/data-analytics-and-algorithms-policingengland-and-wales-towards-new.

[48] For risks attached to using AI in predictive profiling as a measure to prevent crime, see also: K. Blount, *Using artificial intelligence to prevent crime: implications for due process and criminal justice*, «AI & Society», n. 39, pp. 359-368. https://doi.org/10.1007/s00146-022-01513-z.

[49] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, cit.; M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, cit.

[50] See P. Hayes, N. Fitzpatrick, *Narrativity and responsible and transparent AI practices*, «AI & Society», 2024, https://doi.org/10.1007/s00146-024-01881-8. These requirements include participatory and collective intelligence design, taking context into consideration, trustworthiness, accountability, transparency, and legality.

[51] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, cit.

this bias. In any case, we hope we provided additional considerations to motivate using algorithmic processing to support positive social change.

Doing justice to algorithms. Integrating fairness metrics with a structural understanding of justice[a]

Enea Lombardi[*]

Abstract

Questo articolo esplora i limiti dell'equità algoritmica, in particolare il "teorema dell'impossibilità dell'equità", e discute come una comprensione strutturale della giustizia possa affrontare le relative questioni etiche. Dopo aver presentato i principali modelli di equità algoritmica, sostengo che essi trascurano questioni fondamentali di giustizia, dando priorità a metriche basate sui risultati e isolando il processo decisionale da contesti socio-storici più ampi. Inoltre, quando i tassi di base differiscono, diventa impossibile soddisfare contemporaneamente più di una metrica di equità. Per ovviare a queste carenze, propongo di integrare l'equità algoritmica con la nozione di ingiustizia strutturale di Iris M. Young, che tiene conto delle disuguaglianze radicate nell'interazione tra comportamenti, norme e istituzioni. Questo approccio colloca gli algoritmi nel loro contesto socio-storico più ampio, sottolineando i fattori sistemici che influenzano il processo decisionale e perpetuano risultati ingiusti. Sostengo inoltre che una prospettiva strutturale assegna agli algoritmi un duplice ruolo, in particolare nei casi controversi in cui sono in gioco questioni etiche. In primo luogo, una funzione diagnostica: mettendo in luce gli squilibri e i pregiudizi etici sottostanti, gli algoritmi possono evidenziare le aree critiche per le riforme sistemiche. In secondo luogo, possono fungere da strumenti di valutazione, consentendo la valutazione e la definizione delle priorità delle metriche di equità caso per caso.

Parole chiave: equità algoritmica; ingiustizia strutturale; metriche di equità; riforme sistemiche

This paper explores the limitations of algorithmic fairness, particularly the "impossibility theorem of fairness", and discusses how a structural understanding of justice can address the related ethical concerns. After presenting the main models of algorithmic fairness, I argue that they overlook key justice concerns by prioritizing outcome-based metrics and isolating decision-making from broader socio-historical contexts. Furthermore, when base rates differ, it becomes impossible to satisfy more than one fairness metric simultaneously. To address these shortcomings, I propose integrating algorithmic fairness with Iris M. Young's notion of structural injustice, which accounts for entrenched inequalities rooted in the interplay of behaviours, norms, and institutions. This approach situates algorithms within their broader socio-historical context, emphasizing systemic factors that influence

[a] Received on 31/01/2025 and published on 09/12/2025.
[*] Utrecht University, e-mail: e.lombardi@students.uu.nl.

decision-making and perpetuate unjust outcomes. I further contend that a structural perspective assigns algorithms a twofold role, particularly in contentious cases where ethical controversies are at play. First, a diagnostic function: by exposing underlying ethical imbalances and biases, algorithms can highlight critical areas for systemic reforms. Second, they can serve as evaluative tools, enabling the assessment and prioritization of fairness metrics on a case-by-case basis.

Keywords: algorithmic fairness; structural injustice; fairness metrics; systemic reforms

1. Algorithmic Metrics and the Impossibility of Fairness

Algorithms play an increasingly significant role in shaping social and institutional decision-making, particularly in critical areas such as healthcare resource allocation and pretrial risk assessments. As their application expands across various societal domains, concerns have grown regarding their potential to amplify unjust discrimination and biases. These risks are particularly pronounced when it comes to social groups with sensitive attributes, such as race and gender, which are often factors of systemic injustice. For instance, research has shown that facial recognition algorithms are systematically less accurate when applied to non-white women[1]. Additionally, recidivism prediction algorithms reportedly assign higher risk scores to non-white people, particularly in the U.S. criminal justice system[2]. To address these ethical imbalances, many authors have focused on algorithmic fairness[3].

The notion of fairness generally refers to the moral rightness of a given process, particularly in the context of decision-making. Despite differences among various approaches, the underlying intuition is tied to the idea of equality: a procedure is fair if it is equal, meaning it treats people equally in relevant respects[4]. In algorithmic design, fairness has predominantly been operationalized at the group level rather than focusing on individuals. While individual differences are recognized, current approaches seek to mitigate discrimination and biases by addressing fairness at the group scale, benefiting individuals as members of those groups[5]. Among the many existing metrics, the following overview focuses on the most significant and representative models.

First, the disparate impact metric[6] evaluates fairness by requiring that the ratio of positive outcomes between groups exceeds a specific threshold, often 80%. For example,

[1] See J. Buolamwini, T. Gebru, *Gender shades: Intersectional accuracy disparities in commercial gender classification*, in «Conference on fairness, accountability, and transparency», 2018, pp. 77-91.

[2] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, in «ProPublica», 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] See J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, *Algorithmic fairness*, in «AEA Papers and Proceedings», 108, 2018, 22-27; D. Pessach and E. Shmueli, *Algorithmic Fairness*, in L. Rokach, O. Maimon, and E. Shmueli (edited by), *Machine Learning for Data Science Handbook*, Springer, Dordrecht 2023, pp. 867-886; S. Barocas, M. Hardt, A. Narayanan, *Fairness and machine learning: Limitations and opportunities*, MIT Press, Boston 2023.

[4] See J. Broome, *Fairness*, in «Proceedings of the Aristotelian Society», 91, n. 1, 1991, pp. 87-102.

[5] See S. Verma, J. Rubin, *Fairness definitions explained*, in «2018 IEEE/ACM International Workshop on Software Fairness (FairWare)», 2018, pp. 1-7.

[6] See M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, *Certifying and removing disparate impact*, in «Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining», 2015, pp. 259-268.

if a hiring algorithm selects 20% of male applicants but only 12% of female applicants, the resulting disparate impact ratio of 0.6 falls below the 0.8 threshold, indicating potential gender bias. In contrast, demographic parity[7] focuses on equalizing the absolute rates of positive outcomes across groups, regardless of underlying differences and sensitive attributes. To exemplify, this model suggests that the hiring rate for women and men should be exactly the same, even if their average qualifications differ. While both metrics aim to promote equity, they may lead to unintended consequences when base rates differ significantly. The notion of base rate refers to the probability of an event occurring within a specific population before applying any predictive model. In the case of disparate impact, groups with higher base rates could be disadvantaged by suppressing positive outcomes, while demographic parity might disregard meaningful variations such as differences in qualifications or experience. To address these challenges, the strategy of equalized odds[8] proposes a more nuanced fairness criterion by requiring that both the true positive rate and the false positive rate are equal across groups defined by sensitive attributes. This ensures that the model's ability to correctly identify outcomes – and its likelihood of making mistakes – is not biased toward any particular group. For example, if a medical diagnostic algorithm correctly identifies 90% of actual positive cases (true positive rate) and falsely flags 10% of actual negative cases (false positive rate) in both white and Black patients, it satisfies equalized odds. Differently, predictive parity[9] ensures that the positive predictive value – the proportion of predicted positives that are actually correct – is consistent across groups. For instance, if a loan approval algorithm predicts repayment correctly for 80% of approved applicants in both male and female groups, it satisfies predictive parity. Error rate parity[10], on the other hand, focuses solely on balancing specific types of errors (e.g., false positives or false negatives), without requiring both to be equal simultaneously, as equalized odds does. For example, a recidivism risk algorithm that has a false positive rate of 20% and a false negative rate of 10% for both Black and white defendants satisfies error rate parity, even though it may not satisfy equalized odds if only one type of error is balanced.

All these metrics are closely tied to the broader notion of accuracy, which defines the proportion of correct predictions made by an algorithmic model. However, accuracy alone is insufficient to guarantee fairness. For a model with high accuracy may still perpetuate systemic inequalities if it exhibits disparate error rates or ignores sensitive asymmetries in base rates[11]. The difference in base rates is critical in this context, as it implies that some metrics are mutually exclusive, making the achievement of complete fairness impossible. This leads to the "impossibility theorem of fairness", which states that when base rates are different, «it is not possible to satisfy multiple notions of fairness simultaneously»[12]. This is particularly problematic because in real life base rates often differ across groups, and issues of injustice frequently arise precisely due to these differences.

[7] See T. Calders, S. Verwer, *Three naive Bayes approaches for discrimination-free classification*, in «Data Mining and Knowledge Discovery», 21, 2010, 277-292.
[8] See M. Hardt, E. Price, N. Srebro, *Equality of opportunity in supervised learning*, in «Advances in Neural Information Processing Systems», 2016, pp. 3315-3323.
[9] See D. Hellman, *Measuring algorithmic fairness*, in «Virginia Law Review», 106, n. 4, 2020, pp. 811-866.
[10] See J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, *Algorithmic fairness*, cit., pp. 22-27.
[11] See J. Herington, *Measuring Fairness in an Unfair World*, in «Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society», 2020, pp. 286-292.
[12] D. Pessach and E. Shmueli, *Algorithmic Fairness*, cit., p. 873.

Yet, as demonstrated by Chouldechova, «if the base rate […] differs across groups, any instrument that satisfies predictive parity at a given threshold […] must have imbalanced false positive or false negative errors rates at that threshold»[13]. This implies that predictive parity, equalized odds, and demographic parity end up being mutually exclusive for a well-calibrated classifier. For achieving calibration – i.e., ensuring an accurate alignment between predicted probabilities and actual outcomes – contrasts with attaining equality in false negative and false positive rates. This is because groups with different base rates inherently require different thresholds to align probabilities with actual outcomes, but changing thresholds affects the error rates. As a result, ensuring fairness in one aspect will necessarily lead to unfairness in another, as these metrics pull decision thresholds in opposite directions[14].

The impossibility of fairness results in a decisional impasse that offers no baseline for evaluating the moral status of algorithms, particularly in contentious cases – i.e., cases characterized by ethical controversies, such as entrenched inequalities, and contextual variables that are difficult to operationalize. Consequently, this renders some critical ethical concerns unsolvable by fairness metrics alone. These concerns encompass a range of issues, from the case of COMPAS, a pretrial risk assessment algorithm criticized by ProPublica for disadvantaging African American individuals[15], to matters of epistemic hermeneutical injustice in healthcare[16]. As I will show in the next section, the COMPAS case further illustrates that current metrics are inherently incapable of addressing systemic inequalities across groups. To tackle these challenges, I will argue that contentious cases should be addressed by complementing algorithmic fairness with a structural understanding of justice.

2. Algorithmic Fairness and the Structural Understanding of Justice

First proposed by Iris M. Young[17], the notion of structural injustice was developed in contrast to distributive paradigms, which conceive justice as the morally proper allocation of societal goods – framed as resources, opportunities, or welfare – to address unfair social inequalities. While distributive accounts are undoubtedly effective in certain contexts, such as for the allocation of economic resources following an environmental disaster, Young argues that it does not fully capture the scope of justice. To introduce this criticism, she advances a "structural objection", which can be divided into two main claims. First, (i) distributive paradigms overlook the institutional and socio-historical context that shapes the allocation of goods. This oversight results in a twofold harm: it reinforces the underlying context while failing to address the systemic origins of entrenched inequalities. For instance, a fair distribution of economic resources to unemployed people may be beneficial in the short term, but it fails to tackle the systemic factors that necessitated such a distribution in the first place. Consequently, Young argues that (ii) the exclusive focus on distribution neglects the structural dimension of justice. A social structure is an organized

[13] A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, in «Big Data», 5, 2017, pp. 153-163: 158.
[14] See R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, *Fairness in criminal justice risk assessments: The state of the art*, in «Sociological Methods & Research», 50, n. 1, 2021, pp. 3-44.
[15] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, cit.
[16] See G. Pozzi, *Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare*, in «Ethics and Information Technology», 25, n. 3, 2023, pp. 1-12.
[17] See I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990.

field of social positions stemming from «accumulated outcomes of actions of masses of individuals»[18] and institutions acting «according to normally accepted rules and practices»[19]. It represents the skeleton of the social fabric, situating people in positions that are relational, mutable, and characterized by varying degrees of power relations. For example, a specialized working woman may be subjected to the power of her superiors due to her gender, but she can also exert power over unspecialized subordinates. Given that this approach prioritizes systemic constraints, structural injustice occurs when «social processes put large groups of persons under systematic threat of deprivation of the means to develop and exercise their capacities»[20]. Accordingly, since injustice arises from the multifaceted position occupied within the social structure, it cannot be traced back to individual actions or isolated policies. Rather, it emerges from the intertwined relationships among individuals and institutions that are entrenched in everyday practices. This means that structural injustice cannot be rectified solely through ex-post distribution, as the structure functions as a pre-existing field of norms and social positions.

As Kasirzadeh suggests, «most mathematical metrics of algorithmic fairness are inherently rooted in a locally distributive conception of justice», insofar as «they are concerned with how the algorithm would allocate the relevant computational or material goods across different groups»[21]. This resemblance stems from the fact that fairness metrics typically assess how benefits or harms – such as loans, jobs, or medical treatments – are distributed among predefined groups based on sensitive attributes like race or gender. Much like distributive theories of justice, these metrics aim to ensure proportional or equal allocation, without necessarily questioning the structural conditions that shape individuals' positions or access to resources in the first place. Rather, these metrics are predominantly outcome-focused, evaluating fairness at the point of decision-making while abstracting away from the broader social, historical, and institutional contexts that shape both opportunities and data.

As a result, the obstacles faced by algorithmic metrics of fairness are akin to those encountered by distributive paradigms of justice. On closer inspection, the primary obstacles arise from the focus of the normative lenses: the metrics of algorithmic fairness focus solely on current outcomes, overlooking the underlying dynamics that produce them and relying «on a narrow frame of analysis restricted to specific decision points, in isolation from the context of those decisions»[22]. This is exemplified by the case of COMPAS. For the exclusive focus on outcomes has led to conflicts among fairness metrics, which are applied without consideration of the socio-historical context that systematically disadvantages African American individuals. In turn, the implementation of algorithmic metrics results in a situation where «even the best-case scenario – a perfectly accurate risk assessment – would perpetuate racial inequity»[23]. Therefore, since «fairness is operationalized in terms of isolated decision-making processes»[24] and algorithmic metrics

[18] I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, p. 62.
[19] Ivi, p. 100.
[20] Ivi, p. 56.
[21] A. Kasirzadeh, *Algorithmic fairness and structural injustice: Insights from feminist political philosophy*, in «AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society», 2022, pp. 349-356: 351.
[22] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, in «Philosophy & Technology», 35, n. 90, 2022, pp. 1-32: 4.
[23] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, cit., p. 3.
[24] Ivi, p. 4.

are outcome-based, they are inherently incapable of preventing the perpetuation of long-lasting cycles of inequality»[25]. In contrast, a structural understanding of justice combines the evaluation of outcomes with a normative analysis of the context, systemic factors, and social processes that shape these results.

By focusing on the systemic origins of entrenched inequalities, the structural approach introduces three key elements to the normative analysis of algorithms[26]. First, since injustice cannot be fully attributed to individual actors, but emerges from the complex interplay of norms and institutions, decision-making processes must be understood within their broader, systemic contexts. Algorithms, as integral components of these structures, are "sociotechnical entities"[27] that should not be assessed in isolation but must be evaluated as part of the larger fabric of social, economic, and political forces. Second, structural injustice is forward-looking. Rather than simply focusing on offering reparations for past wrongs, this approach prioritizes reforms to current and future conditions. It calls for proactive measures that seek to address the causes of entrenched inequalities by detecting and implementing structural conditions that can guide the decision-making process toward more equitable outcomes. Third, structural injustice is not a static notion. Since structures can accumulate and compound over time, the temporal dynamics of injustice – i.e., how it unfolds and deepens across time – are crucial to the normative analysis of algorithms. This highlights the importance of continuously auditing and reassessing datasets, which are never entirely neutral or value-free. For datasets are shaped by historical and contextual factors that can encode biases and reproduce inequalities, thereby placing algorithms trained on them at risk of perpetuating the very injustices they aim to mitigate[28].

These elements suggest the need for a structural approach that incorporates systemic and contextual considerations into the analysis of algorithmic decision-making. According to Green[29], this model involves three main steps. The first is identifying inequalities by detecting entrenched disparities and examining how social, political, and institutional arrangements reinforce them. This involves analyzing power structures and understanding how deep-rooted practices contribute to systemic inequities. The second stage focuses on enacting reforms, determining what changes could mitigate the identified inequalities, and emphasizing the need to restructure decision-making processes to reduce their role in exacerbating social hierarchies. This phase is not just about implementing technical fixes, but about transforming the systems and practices that give rise to inequality. The third step involves assessing the role of algorithms by analyzing whether and how they can support these reforms. While algorithms can be integral to certain aspects of systemic change, they should be viewed not as standalone solutions, but as contextual tools operating within broader reform agendas. Therefore, this approach suggests that a structural understanding of justice can address the impossibility of fairness in three ways, particularly in contentious cases where ethical controversies are at play.

---

[25] A. Kasirzadeh, *Algorithmic fairness and structural injustice: Insights from feminist political philosophy*, cit., p. 352.

[26] Cfr. J. Himmelreich, D. Lim, *AI and structural injustice*, in J.B. Bullock, Y. Chen, J. Himmelreich, V.M. Hudson, A. Korinek, M.M. Young, B. Zhang (edited by), *The Oxford Handbook of AI Governance*, Oxford University Press, Oxford 2022, pp. 210-231.

[27] See A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, *Fairness and abstraction in sociotechnical systems*, in «Proceedings of the Conference on Fairness, Accountability, and Transparency», 2019, pp. 59-68.

[28] See C. Stinson, *Algorithms are not neutral*, in «AI Ethics», 2, 2022, pp. 763-770.

[29] See B. Green, *Escaping the impossibility of fairness*, cit.

First, algorithms can function more as diagnostic tools that reveal systemic imbalances rather than offering direct solutions. Consider the case of COMPAS, which can be viewed as an instance of structural injustice[30]. Rather than being used to solve pretrial risk assessment, it can be assigned an *ex-negativo* role. This implies that its primary task is not to provide prescriptive guidance on what should be done – such as extending incarceration – but rather to highlight the ethical controversies, biases, and technical ambiguities involved, which can help refine the focus of reform agendas. Accordingly, COMPAS might be employed to expose systemic discrimination against African American people and illuminate potential areas for targeted reforms. To illustrate, this could involve addressing criminogenic conditions in disadvantaged communities, such as through the redesign of urban environments[31] or reforms to the educational system[32], to amend the social structure and reduce systemic recidivism from the outset.

Second, as tackling entrenched injustice is a long-term and multifaceted process, this structural approach can also serve as an evaluative tool to assess and prioritize fairness metrics on a case-by-case basis. Rather than dismissing algorithmic metrics altogether, the structural approach allows for a context-sensitive deployment of such tools by treating fairness metrics not as universally applicable standards, but as evaluative instruments whose relevance must be determined in light of the specific injustices at play. This case-by-case orientation acknowledges that no single metric can capture the full scope of fairness in every situation, but that certain metrics – when interpreted through a structural lens – can help reveal, monitor, and eventually mitigate the particular forms of inequality embedded in distinct socio-institutional contexts. To illustrate, consider again the case of COMPAS. While it failed to satisfy the metric of error rate parity, since African American non recidivists were more likely to be classified as high risk[33], it complied with predictive parity, as outcomes were predicted at the same rate across all groups[34]. In this framework, the structural approach would suggest assessing the two metrics from a systemic perspective. This involves evaluating the algorithm within the broader socio-historical context and determining whether it mitigates or exacerbates existing biases and discrimination. Accordingly, error rate parity performs better from a structural standpoint because, while not exhaustive, its outcomes align with the systemic discrimination experienced by African American people and are responsive to the demands of the social context. This suggests that, when faced with the impossibility of fairness, the structural approach can overcome the decisional impasse by introducing an additional normative layer that broadens the locus of assessment and selects one metric over another. This can be viewed either as a transitional step toward achieving perfectly just outcomes, or as a permanent stage in which fairness is continuously evaluated and refined by justice.

Importantly, this perspective acknowledges that datasets will never be neutral, as they are shaped by historical power asymmetries, institutional practices, and societal

[30] See A. Kasirzadeh, *Algorithmic fairness and structural injustice*, cit.

[31] See P.M. Cozens, *Sustainable urban development and crime prevention through environmental design for the British city: Towards an effective urban environmentalism for the 21st century*, in «Cities», 19, n. 2, 2002, pp. 129-137.

[32] See L. Lochner, E. Moretti, *The effect of education on crime: evidence from prison inmates, arrests, and self-reports*, in «American Economic Review», 94, n. 1, 2004, pp. 155-189.

[33] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, cit.

[34] See A.W. Flores, K. Bechtel, C.T. Lowenkamp, *False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks"*, in «Federal Probation», 80, 2016, pp. 38-46.

inequalities. Yet, by embedding fairness metrics within a structural evaluative lens, such biases can be rendered visible, interrogated, and, to some extent, mitigated – turning algorithmic evaluation into a site of critical reflection and corrective intervention. To illustrate, the structural approach is particularly well-equipped to account for the epistemic risks posed by confirmation bias and self-fulfilling prophecies. Rather than accepting algorithmic outputs as neutral or dispositive, this framework embeds them within a recursive process of critical scrutiny. Tools such as COMPAS are not interpreted as delivering authoritative verdicts, but as indicators of deeper systemic dynamics that warrant interrogation. By assigning algorithms an *ex-negativo*, diagnostic function, the structural approach resists the closure of interpretive loops that confirmation bias typically exploits. Instead of allowing predictive scores to validate entrenched assumptions – such as the presumed higher risk of certain demographic groups – the algorithm's role is to surface and destabilize such associations, thereby revealing the socio-political structures that generate them. In this sense, algorithmic metrics serve not to confirm pre-existing beliefs, but to provoke epistemic friction and redirect inquiry toward underlying institutional and historical causes. Thus, rather than erasing confirmation bias, the structural approach renders it visible and accountable, integrating its acknowledgment into the broader pursuit of justice.

Third, this approach promotes a forward-looking, shared responsibility[35] among all stakeholders, including the developers and the implementers of algorithms. This implies that responsibility does not rest solely with manufacturers but extends to users – such as judges in the case of COMPAS – who cannot disregard ethical contentions due to the involvement of multiple actors. Rather, they must adopt a responsible and transparent approach to the use of algorithms, recognizing that they are active participants who could contribute to exacerbating systemic inequalities[36]. This entails that since the impossibility of fairness implicates all actors, conflicting fairness metrics require stakeholders to reassess datasets and contexts in line with their roles within the broader structure. In the case of COMPAS, this could involve judges establishing a committee to evaluate ethical concerns, identify risks, clarify ambiguities and biases – particularly in datasets – and prevent potential misuses. Ultimately, this underscores a further point of distinction of the structural approach: responsibility is a critical component of justice, as it represents the first step toward forward-looking reforms.

One might object that this approach underestimates the role of algorithmic metrics, rendering the notion of fairness redundant and normatively useless. It might seem that normative work is solely carried out by the structural approach, which ends up overshadowing the importance of current metrics. I reply in three ways. First, this approach is designed to complement, not replace, fairness metrics. This is for two reasons. On the one hand, a structural approach is much more complex to operationalize compared to the notion of fairness, as the variables involved are contextual, non-linear, and interrelated[37]. On the other hand, it acknowledges that there are cases where fairness metrics are sufficient – for instance, when base rates are significantly similar, and the context is not characterized by systemic inequalities. Second, although this approach emphasizes the need for human

[35] See I.M. Young, *Responsibility for Justice*, cit.

[36] See R.E. Goodin, C. Barry, *Responsibility for structural injustice: A third thought*, in «Politics, Philosophy & Economics», 20, n. 4, 2021, pp. 339–356.

[37] See B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, cit.

supervision of the decision-making process, it does not stand in contrast to algorithmic metrics per se. Rather, it advocates for the design, implementation, and cross-use of different algorithms to incorporate systemic and contextual factors, while emphasizing the need for ongoing human oversight and assessment of both the datasets and the results. Third, this approach – as Young[38] emphasizes – moves from injustice rather than justice. In this context, this means that while it aims to design increasingly just algorithms, its normative task arises from the shortcomings of algorithmic metrics and the impossibility of achieving complete fairness. This implies that the operationalization of fairness remains essential, as the structural approach builds upon the results of these metrics – ultimately recognizing their normative validity and utility, as demonstrated by the prioritization of one metric over another. Therefore, this approach acknowledges both the limitations and potential of algorithmic metrics while emphasizing the necessity of ongoing human oversight, ethical reflection, and collective responsibility in designing and implementing fair decision-making processes and just outcomes.

One might further question the use of algorithms altogether. Given their entrenched biases and their *ex-negativo* role, it could be argued that the use of algorithmic metrics seems unjustified from the outset. I reply that, even when the ultimate goal is to transform upstream decision-making, quantitative diagnostics remain indispensable. Without such metrics, systemic injustices embedded in data sets might remain hidden, lacking the provisional benchmarks needed to make them visible, comparable, and manageable. Although fairness metrics often identify injustice only after it has occurred, this retrospective insight is vital, as it uncovers entrenched biases that would otherwise remain undetectable from a quantitative perspective. While this is surely a limitation inherent to algorithms, it simultaneously underscores a key strength of the structural approach: it does not passively wait for injustice to unfold, but rather integrates algorithmic metrics within a broader agenda of anticipatory and preventive interventions. By embedding these tools in a framework oriented toward systemic reform – through policies, institutional redesign, and cultural transformation – the structural approach ensures that the detection of bias is only one component in a wider strategy aimed at dismantling its root causes before they materialize in concrete harms. Properly contextualized, then, algorithmic tools can function as provisional instruments that "mathematize" injustice, offering a quantitative baseline from which to evaluate discrimination and enable structured comparisons across cases and contexts. In turn, these metrics serve an educative role, illuminating the persistence and extent of biases, thereby anchoring the political and institutional will necessary for upstream reform. In this way, algorithmic evaluation does not replace the imperative for structural transformation but reinforces it, providing the empirical foundation essential to dismantling injustice at its roots.

3. Conclusion

This paper explored how Iris M. Young's structural understanding of justice can address the limitations of algorithmic fairness metrics, in particular the "impossibility theorem of fairness". I began by demonstrating that operationalizing fairness in algorithms often

[38] See I.M. Young, *Justice and the Politics of Difference*, cit.

overlooks critical ethical concerns, as it is primarily outcome-based and isolates decision-making from its broader socio-historical context. As a result, systemic injustices risk being neglected, embedded within datasets as neutral information, and eventually reinforced by algorithmic decision-making. I further referred to the "impossibility theorem of fairness" to illustrate that when base rates differ, current metrics are mutually exclusive, rendering the achievement of full fairness unattainable. To tackle these challenges, I proposed adopting a structural understanding of justice to contextualize algorithms and facilitate systemic assessment. I argued that, from a structural perspective, algorithms can serve as (i) diagnostic tools that reveal systemic inequalities and ethical imbalances, thereby identifying critical areas for forward-looking reforms, and (ii) evaluative tools that, in contentious cases, assess and prioritize fairness metrics on a case-by-case basis. Furthermore, I emphasized that this approach fosters a shared account of responsibility among diverse stakeholders, including developers, users, and decision-makers. In conclusion, a structural understanding of justice can overcome the impossibility of fairness by evaluating the normative implications of algorithms and the value-laden nature of datasets, contextualizing decision-making processes, distributing responsibility among all stakeholders, and assuming a systemic perspective that can assess and rank fairness metrics tailored to each case. While this paper focused on contentious cases, a broader caution would suggest integrating the structural approach in all cases, as it consistently offers prescriptive guidance to complement, oversee, and enhance algorithmic fairness across various domains.

A Kantian approach to the possibility of mechanical consciousness[a]

Sung-Yeop Jo[*]

Abstract

Il progresso dell'intelligenza artificiale (IA) mette in discussione le definizioni tradizionali di coscienza. Questo articolo applica la filosofia di Immanuel Kant per sostenere che, sebbene la capacità di un'IA di sintetizzare i dati sensoriali in rappresentazioni oggettive soddisfi il criterio funzionale di una coscienza rudimentale, come delineato nella sua Deduzione A, ciò deve essere distinto dalla coscienza di sé più robusta della Deduzione B. Questa consapevolezza di ordine superiore, identificata con il "penso" o l'appercezione originale, è un atto spontaneo di un sé unificato e la base per un vero e proprio agire. Kant fonda i diritti morali e legali sullo status di persona, che richiede tale agire autonomo e autocosciente. Essendo un sistema che segue esclusivamente delle regole, un'IA non ha la capacità di formulare le proprie massime da una prospettiva in prima persona. Pertanto, sebbene un'IA possa essere considerata "cosciente" in senso kantiano limitato, non può qualificarsi per lo status etico o giuridico di persona.

Parole Chiave: intelligenza artificiale, coscienza, appercezione, persona, agire

The advancement of artificial intelligence (AI) challenges traditional definitions of consciousness. This article applies the philosophy of Immanuel Kant to argue that, while an AI's capacity to synthesize sensory data into objective representations meets the functional criterion for a rudimentary consciousness as outlined in his A-Deduction, this must be distinguished from the more robust self-consciousness of the B-Deduction. This higher-order awareness, identified with the 'I think' or original apperception, is a spontaneous act of a unified self and the basis for genuine agency. Kant grounds moral and legal rights in the status of personhood, which requires this autonomous, self-conscious agency. As a purely rule-following system, an AI lacks the capacity to formulate its own maxims from a first-person perspective. Therefore, while an AI may be considered 'conscious' in a limited Kantian sense, it does not qualify for the ethical or legal status of a person.

Keywords: artificial intelligence, consciousness, apperception, person, agency

[a] Received on 28/03/2025 and published on 09/12/2025.

[*] Ludwig-Maximilians-Universität München, Chair of Metaphysics and Ontology, e-mail: sung-yeop.jo@outlook.com.

1. Introduction

Certain beings are typically regarded as conscious, yet it is still far from obvious what consciousness itself is or how the term ought to be defined. This long-standing philosophical challenge has acquired new urgency with the rise of artificial intelligence (AI) models that exhibit human-level cognitive capacities. Indeed, given the variety and quality of their outputs, asserting a fundamental difference between their intelligence and our own has become increasingly difficult.

The central task of this article, therefore, is to employ the framework of Kantian philosophy to clarify the essential conditions of human consciousness and subsequently determine whether these same conditions could, at least in principle, apply to AI. To address this issue, I will return to Kant's *Critique of Pure Reason* (CPR), which offers a profound analysis of the necessary conditions for consciousness. My focus will be the A- and B-Deductions, where Kant defends the actuality of human consciousness and its necessary unity against the empiricist critics, most notably David Hume. In response to Hume's bundle theory, which reduces the thinking self to a mere collection of fluctuating perceptions, Kant shifts focus to the mental act of synthesis, arguing that it is through this original act that disparate sense perceptions are combined into a unified, objective representation. The necessary unity of consciousness for objective representation – the principle of apperception (A117) – constitutes the very foundation of the Kantian theory of consciousness.

The challenge, however, is that the Kantian principle of apperception also seems applicable to advanced AI. An AI model that can cognize objects in a visible image, for instance, seems to perform the very synthesis of sensory data into objective representations that Kant describes. This raises two crucial questions: First, if we accept the premises of Kant's philosophy of consciousness, on what grounds could we then deny that an AI with this capability possesses a form of consciousness? Second, if we were to grant it consciousness, would it follow that such an entity also deserves ethical consideration?

To develop the full potential of the Kantian approach to the possibility of mechanical consciousness, this paper will proceed in two main parts. First, drawing on a selective reading of Kant's A-Deduction, I will establish that the Kantian criterion for consciousness is the mental ability to identify empirical objects by applying concepts to sense perceptions. I will then argue that advanced AI models that meet this criterion should be recognized as possessing a rudimentary form of consciousness. Second, I will sharpen and defend this claim by clarifying that this rudimentary consciousness is conceptually distinct from the more robust *self-consciousness* that Kant identifies as a necessary condition for genuine agency. This distinction explains why such a "conscious" AI would not have a claim to human-like rights, such as those for their intellectual contributions to research. I will develop this crucial difference between consciousness and self-consciousness by highlighting the specific developments in Kant's argument from the A- to the B-Deduction.

2. Kant's Concept of Consciousness in the A-Deduction

I. Synthesis as the Ground of Consciousness

If consciousness is a capability of the mind, it must manifest itself through mental activities. The crucial question, however, is what specific activities serve as this manifestation and how they can provide reliable evidence for the actuality of consciousness. In the A-Deduction, Kant provides a clear answer to this inquiry, writing as follows:

> Without consciousness that which we think is the very same as what we thought a moment before, all reproduction in the series of representations would be in vain. For it would be a new representation in our current state, which would not belong at all to the act through which it had been gradually generated, and its manifold would never constitute a whole, since it would lack the unity that only consciousness can obtain for it. […]
> For it is this *one* consciousness that unifies the manifold that has been successively intuited, and then also reproduced, into one representation. This consciousness may often only be weak, so that we connect it with the generation of the representation only in the effect, but not in the act itself, i.e., immediately; but regardless of these differences one consciousness must always be found, even if it lacks conspicuous clarity, and without that concepts, and with them cognition of objects, would be entirely impossible (A103; emphasis in original)[1].

These two passages mark the first time in the A-Deduction that Kant uses the term 'consciousness'. Here, he continues his analysis of objective cognition by examining the contribution of concepts to synthetic judgment *a priori*, building upon his preceding discussions of the synthesis of apprehension in intuition and the synthesis of reproduction in imagination. The necessary contribution of consciousness to cognition, Kant argues, is to produce a mental representation of an empirical object by applying a concept to a manifold of sense perceptions, which has already been provisionally combined by the prior syntheses of apprehension and reproduction. In the B-Deduction, Kant employs the example of cognizing a weighty body to illustrate this synthetic act of our mind. When one holds a body – such as a stone – and feels its downward pressure, one forms the judgment that *it is heavy*. However, as Kant points out, the raw sensations alone would amount to nothing more than a subjective sequence of perceptions, such as «If I carry a body, I feel a pressure of weight». Arriving at the objective representation, «It, the body, *is* heavy», requires an additional mental act to combine these underlying sense perceptions into a single, united representation (B142; emphasis in original). This demonstrates, for Kant, that forming an objective representation necessitates a synthetic combination of disparate perceptions – an act performed by consciousness.

The cautious nature of Kant's inference to the principle of apperception bears repeating. He does not directly address the presence of consciousness; rather, he infers its necessary unity from the given fact that a manifold of sense perceptions is *found combined* in individual representations of our mind. Instead of attempting to access consciousness firsthand, which Kant regards as too "weak" to be an object of distinct reflection, he posits the principle of apperception as a necessary condition for objective representation, taking the reality of that representation as his argument's starting point. This marks an

[1] I. Kant, *Critique of Pure Reason*, in *The Cambridge Edition of the Works of Immanuel Kant*, edited by P. Guyer and A.W. Wood, Cambridge University Press, Cambridge 1998.

argumentative strength of the Kantian approach to consciousness. Unlike the attempts to ground consciousness in empirical observation via inner sense – a method whose reliability is philosophically dubious – Kant's principle of apperception is not an empirical finding. Rather, it is the result of a transcendental inference to the necessary conditions for our objective representation. It is the undeniable actuality of the empirical representations of objects that justifies the positing of this 'weak' consciousness and its synthetic function.

In Kant's *Deduction*, consciousness is therefore presented not as a passive state, but as the act of synthesizing disparate sense perceptions into a single, conceptual representation. This active nature of consciousness is revealed in the application of *concepts* during the act of synthesis. As the preceding example in the B-Deduction implies, the *a priori* combination of sense data imparts objectivity to our representations. Their objectivity consists in the crucial distinction between a merely subjective, contingent association, where perceptions are simply simultaneous in the mind, and an objective synthesis, where those same perceptions are represented as necessarily combined in the object itself. The concept of a body, for instance, entails diverse properties, such as extension and mass. Therefore, by applying this concept, a conscious subject is able to unite the distinct sense perceptions of a certain extension and a certain mass into the single representation of one object. In performing «the synthesis of recognition in the concept» (A103), consciousness thus reveals itself as the faculty that makes objective recognition possible. It is through this specific synthetic act, which uses a concept as its rule, that an object can be empirically cognized[2].

The juxtaposition of Kant's inference to the necessary unity of consciousness in the A-Deduction with recent developments in AI technology leads us to a crucial philosophical question. To the extent that an AI is able to combine sensory data to identify objects, this ability seems *functionally similar* to the synthesis that Kant described. If we consistently apply Kant's line of reasoning, are we not then compelled to recognize this same capability in a non-human system as another form of consciousness?[3] Indeed, Kant's concept of consciousness in the A-Deduction provides a purely functional criterion for consciousness. On this view, consciousness is defined not by its biological substrate or private feeling but by the specific act of forming an objective representation through the application of concepts to sensory data. Therefore, any system that demonstrably performs this synthetic function – whether human or artificial – meets the condition for this rudimentary form of consciousness as established by the A-Deduction's logic.

3. Various Forms of Consciousness in Kant's Philosophy

I. The Logical Status of the Synthetic Act in the A-Deduction: Necessary or Sufficient for Consciousness?

[2] A significant point of scholarly debate regarding the A-Deduction is that Kant does not explicitly define what he means by "concept" in this context. This ambiguity has led to conflicting interpretations. For instance, Guyer argues that Kant is referring exclusively to the pure concepts of understanding, i.e. the categories. In contrast, Kitcher maintains that the argument also applies to, and even requires, empirical concepts, such as dog. See: P. Guyer, *Kant and the Claims of Knowledge*, Cambridge University. Press, Cambridge 1987, pp. 149-154; P. Kitcher, *Kant's Thinker*, Oxford University Press, New York 2011, pp. 126-142.

[3] This conceptual inquiry does not entail a technological evaluation of current AI models. The question of what conditions must be met for any being to be recognized as conscious remains philosophically valid, regardless of the actual capabilities of today's technology.

Before this 'Kantian' approach to the possibility of mechanical consciousness can be fully persuasive, several fundamental challenges must first be met. The first challenge is whether the mental act of combining a manifold of sense perceptions into an objective representation is a *sufficient* condition for attributing consciousness or merely a *necessary* one.

From a purely logical point of view, Kant's transcendental argument for the necessary unity of consciousness is sound if and only if the mental capability of synthetic combination is the sufficient condition for consciousness. In the A-Deduction, consciousness is described as *nothing more than* the functional foundation for the synthetic unity of sense perceptions as these are found combined in objective representations of our mind. Thus, Kant defines the transcendental apperception as «pure, original, unchanging consciousness», without which «no cognitions can occur in us, no connection and unity among them» (A107). The transcendental apperception functions as the foundation for all other forms of self-knowledge, which are acquired through empirical reflection on the contents of inner sense. For Kant, internal reflection is fundamentally empirical, as it is always situated in time and is therefore incapable of validating an *objective* representation of the self, in contrast to the *pure* consciousness of the transcendental apperception. Kant's transcendental apperception is thus "original" in a specific sense: it functions as the foundational premise for all subsequent philosophical discussions of consciousness and self-knowledge. The inference from the combination of sense perceptions in an objective representation to the transcendental apperception that makes it possible, therefore, provides the *only* viable ground for inferring the actuality of this "weak" consciousness, which cannot be confirmed through direct observation but only indirectly through a philosophical reflection on its *effects* – that is, the very representations unified by its synthetic act (A104). That is to say, the argument in Kant's A-Deduction arguably provides a functional criterion for consciousness. It suggests that a subject can recognize herself as conscious *insofar as* she finds objective representations in her mind – representations that necessarily consist of the combination of diverse sense perceptions. In the A-Deduction, Kant's argument attempts to derive the actuality of a unifying consciousness from the very concept of objective combination. For this argument to be sound, the capability for synthetic combination cannot merely be a *necessary* condition for consciousness but must be treated as a *sufficient* one. This is because if this capability were merely necessary, one could not validly infer the actuality of consciousness from the fact of combination *alone*.

In sum, Kant's A-Deduction establishes a sufficient condition for recognizing an entity as conscious: an entity qualifies as conscious *if* it possesses the intellectual capability to combine sense perceptions into a unified representation of an object. This criterion is the result of Kant's strategy for countering Humean skepticism about the self. Instead of searching for the self by means of inner sense, Kant asserts its necessary actuality solely based on the transcendental argument: for a manifold of sense perceptions to constitute an objective representation, they must first be combined by a single, unified consciousness.

II. Applying Kant's Concept of Consciousness to Non-Human Beings

The second challenge is whether Kant's analysis, which takes the form of a first-person reflection on consciousness, can be legitimately applied to other, non-human entities. A

starting point for this inquiry can be found in Kant's *Jäsche Logic*, which provides an illuminating description of the diverse levels of cognitive capacities.

> The *first* degree of cognition is: *to represent* something;
> The *second*: to represent something with consciousness, or *to perceive* (*percipere*)
> The *third*: *to be acquainted* with something (*noscere*), or to represent something in comparison with other things, both as to sameness and as to *difference*;
> The *fourth*: to be acquainted with something with *consciousness*, i.e., to *cognize* it (*cognoscere*). Animals are acquainted with objects too, but they do not *cognize* them.
> The *fifth*: to *understand* something (*intelligere*), i.e. to cognize something through *the understanding by means of concepts*, or *to convince*. One can conceive much, although one cannot comprehend it, e.g. a *perpetuum mobile*, whose impossibility is shown in mechanics.
> The *sixth*: to cognize something through reason, or *to have insight* into it (*perspicere*). With few things do we get this far, and our cognitions become fewer and fewer in number the more that we seek to perfect them as to content.
> The *seventh*, *finally*: to comprehend something (*comprehendere*), i.e., to cognize something through reason or a priori to the degree that is sufficient for our purpose. […]" (Jäsche Logic, IX:64-5; italics in original)[4].

In this passage, Kant draws a notable distinction between *noscere* (to be acquainted with something), a capability he also attributes to animals, and *cognoscere* (to be acquainted with something *with consciousness*). Since a basic consciousness of an object is already presupposed in mere acquaintance, the consciousness that elevates this to genuine cognition must, arguably, be of a different order. In the CPR, Kant defines cognition (*cognitio*) as an objective perception—this is a form of knowledge that is either intuitive or conceptual (CPR A320/B376–7). While animal consciousness can, therefore, discern sensible objects by comparing their similarities and differences, it lacks the ability to (re-)cognize them as *objective* beings, distinct from its own *subjective* awareness[5]. The «synthetic recognition in a concept», which Kant analyzes in the third section of the A-Deduction, corresponds to what he calls *intelligere* in the *Jäsche Logic*. For Kant, human knowledge is discursive: it always requires concepts to synthesize the manifold of sense perception (CPR A68/B93). Human consciousness can not only cognize objects (*cognoscere*) but also understand them (*intelligere*) through its conceptual capacity.

Kant's analysis of various forms of consciousness makes it clear that he accepted the possibility of non-human consciousness such as that of animals. The crucial limitation of animal consciousness is that it lacks a sense of objectivity. In human understanding, by contrast, this objectivity is guaranteed by the application of concepts, as was already demonstrated in the earlier comparison between synthetic combination and mere association. This, in turn, allows us to address the possibility of mechanical consciousness using Kantian framework, since such a system also appears to discern sensible objects using the same kinds of concepts that human beings do[6]. That is to say, if an advanced AI model

[4] I. Kant, *Jäsche Logic*, in *The Cambridge Edition of the Works of Immanuel Kant*, edited by P. Guyer and A.W. Wood, Cambridge University Press, Cambridge 1992.

[5] For a discussion of Kant's illustration of animal consciousness with the example of an ox which can distinguish a stall from a door: P. Kitcher, *Kant's Thinker*, cit., pp.115-119.

[6] In his article examining the possibility of a computer system that satisfies the Kantian conditions for consciousness, Evans also emphasizes the crucial role that concepts play. For Evans, as for Kant, it is the application of concepts that combines disparate sense perceptions into a single, unified reservation, see R.

can cognize sensible objects by applying concepts, such as the concept of body, then on a consistent application of Kantian terminology, it must be recognized as conscious, since conceptual understanding (*intelligere*) necessarily presupposes consciousness.

III. Consciousness versus Self-Consciousness: A Critical Limitation

Not every conscious being is necessarily self-conscious. This crucial difference, which is not sufficiently explicated by Kant himself, requires careful deliberation, particularly when considering the possibility of non-human consciousness. The unique ethical position of human beings is fundamentally grounded in their status as persons. This status, in turn, is typically attributed to their self-consciousness – a mental capacity that animals are considered to lack. It follows, therefore, that if an AI model were regarded as self-conscious, it too would be eligible for a similar ethical status.

According to Kant's analysis of self-consciousness in the B-Deduction, while the synthetic combination of sense perceptions may be a sufficient condition for consciousness in a basic Kantian sense, the mere ability to (re-)cognize objects through such a combination is not, in itself, a sufficient ground for classifying AI as self-conscious. Kant's B-Deduction takes the concept of self-consciousness as its starting point, whereas the A-Deduction only arrives at this concept as its conclusion. In the second section of the B-Deduction (§ 16), Kant introduces his famous principle that the pure representation 'I think' must be able to accompany all of one's representations. He names this reflective awareness of a constant, unified self "original apperception", a form of *self-consciousness* that he defines at length[7]. The project of the B-Deduction, then, is to analyze the fundamental structure of this self-consciousness by examining the principle of the necessary accompaniment of the 'I think' and its philosophical implications. This represents a significant departure from the A-Deduction, in which the analysis does not necessarily presuppose that the cognizing self is self-conscious. In this notable move at the beginning of the B-Deduction, Kant explicitly identifies the agent of the synthetic combination with the first-person pronoun 'I,' whereas in the A-Deduction, such a self is at first only an implication of the transcendental analysis of objective cognition, emerging explicitly only toward the argument's conclusion[8].

However, possessing a certain mental capacity and ascribing its results to oneself are two distinct acts, as the latter is not necessarily entailed by the former. It is arguably this very distinction that makes Kant's cautious, step-by-step approach in the A-Deduction plausible. This, in turn, requires us to distinguish between the 'mere' consciousness and

Evans, *The Apperception Engine*, in H. Kim, D. Schönecker (eds.), *Kant and Artificial Intelligence*, Walter de Gruyter GmbH, Berlin-Boston 2022, p. 67.

[7] «I call it the *pure apperception*, in order to distinguish it from the *empirical* one, or also the *original apperception*, since it is that self-consciousness which, because it produces the representation *I think*, which must be able to accompany all others and which in all consciousness is one and the same, cannot be accompanied by any further representation» (B132; emphasis in original).

[8] Kitcher's six-step reconstruction of the A-Deduction's argument offers a precise analysis of Kant's strategy. According to her reading, the fourth stage of the argument establishes that one becomes aware of representational states and their necessary connections. It is only after this, in the fifth stage, that the thinking self is introduced. Kitcher formulates this step as follows: «When representational states are recognized as necessarily connected together, they are recognized as instances of the I-rule». See P. Kitcher, *Kant's Thinker*, cit., pp. 141-142.

self-consciousness. The former serves simply as a necessary condition for our empirical cognition and does not, in itself, involve a first-person perspective, since the synthetic combination of perceptions into an objective representation implies only that the mental act must be ascribed to a certain conscious subject; it does not entail that the subject is aware of itself performing that act. The latter, by contrast, is expressed in the principle that the 'I think' that must be able to accompany all of one's representations. The principle of the B-Deduction, therefore, necessarily presupposes an additional act of self-reflection: the recognition that all one's representations must be regarded as one's own. This mental act of self-ascription is purely *spontaneous*. The pure awareness of the 'I' is not something given by the senses; rather, the subject attains self-consciousness by reflecting on formal conditions of unity within her own representations[9]. In the B-Deduction Kant proceeds directly from his insight into the necessary accompaniment of the 'I think,' which allows him to address the concept of self-consciousness. This stands in sharp contrast to his more cautious, multistep approach in the A-Deduction, which begins with an analysis of empirical cognition and only derives self-consciousness as a final conclusion.

While Kant's transition to self-consciousness may seem abrupt, it becomes understandable given that the discussions in the A- and B-Deductions are always about consciousness from a first-person perspective. The appeal of this move stems from a subtle phenomenological fact: as a subject, *I* experience an intuitive awareness of a thinking self that accompanies all the objects of my thought. In the B-Deduction, Kant advances to this more robust form of self-consciousness based on a new starting point that the thinking 'I' must always be presupposed as the subject to which all of one's representations belong. The essence of Kantian self-consciousness is this original act of self-ascription: the awareness that all of one's diverse representations belong to a single, unified 'I.' Such an awareness of the 'mineness' of one's representations is a subtle yet constant feature of cognition, which one can bring into focus through an act of reflection.

Kant clarifies the immediate nature of self-ascription in his critique of the metaphysical doctrine of the soul, developed in the *Transcendental Dialectic*. There, he maintains that this mental act of self-ascription is not a logical inference but is, rather, the direct expression of cognition's innate, formal structure, which is grasped through the purely spontaneous reflection on one's own mind. In other words, transcendental apperception does not refer to a substantial self that can be observed or proven to exist; rather, it signifies the transcendental necessity that all of one's representations belong to a single 'I think.' Kant himself describes this principle in the following passage:

> This *I* would have to be an intuition, which, since it would be presupposed in all thinking in general (prior to all experience), would, as an intuition, supply *a priori* synthetic propositions if it were to be possible to bring about a pure rational cognition of the nature of a thinking being in general. Yet this I is no more an intuition than it is a concept of any object; rather, it is the mere form of consciousness, which accompanies both sorts of representations and which can elevate them to cognitions only insofar as something else is given in intuition (A382; emphasis in original)[10].

[9] Strawson makes a similar point, identifying «the necessary unity of consciousness» with «the possibility of self-consciousness». See: P. F. Strawson, *The Bounds of Sense*, Routledge, London-New York 1966, pp. 9-10.
[10] Kant makes this same point in the second edition of Paralogism, arguing that the I of 'I think' is not an object of intuition, but is, instead, the formal expression of self-consciousness (B412).

Notably, the act of self-ascription is only possible for representations contained within one's own mind: one cannot assert that another person thinks in the same way that one does. This is because while an individual ascribes thoughts to oneself through a direct, first-person awareness, one can only ever infer the thoughts of others from a third-person perspective. A subject ascribes one's own thoughts based on a direct and pure awareness of own mental acts—an awareness that constitutes self-consciousness, which Kant's B-Deduction takes as its starting point. One has no such immediate access, however, to the mental states of others; their thoughts can only be inferred, never known. Following the argument of Kant's B-Deduction that the awareness of the 'I think' is a non-inferential and uniquely first-person act, it is therefore impossible to derive any direct knowledge of self-consciousness in other beings.

We can now return to our initial question: Does an AI model qualify as self-conscious if it is capable of (re-)cognizing objects by combining sensory data with concepts? Based on the Kantian distinction between consciousness and self-consciousness, the answer is a clear no. The synthetic capability of our mind, while perhaps a sufficient condition for consciousness, is not in itself sufficient to justify the attribution of self-consciousness. That is to say, one cannot know, in principle, whether an AI system experiences the outputs it produces as 'its own.' For Kant, however, self-consciousness is defined by this very capacity: the original and spontaneous recognition that one's representations are indeed one's own[11].

The sharp distinction between consciousness and self-consciousness is crucial, because it provides a principled basis for resisting the conclusion that AI systems should be granted moral and legal rights. In Kant's practical philosophy, such rights are granted only to beings who qualify as *persons*. In his *Religion within the Boundaries of Mere Reason*, Kant defines a person as a rational being who can be held responsible for his or her actions (Rel VI:26). This responsibility, in turn, arises from the mental capacity to act according to a subjective rule, or *maxim*, that the individual gives to themselves. Kant's categorical imperative thus requires a rational agent to consider whether the maxim grounding their action could be willed as a universal law. The very act of willing a maxim to become a universal law, however, necessarily presupposes a first-person perspective, which constitutes the basis for personal responsibility. Therefore, an AI cannot be considered a genuine moral agent, because its actions are based on pre-programmed rules given to it externally, not on maxims it gives to itself. Furthermore, it lacks the capacity to formulate its own maxims, a process that requires a uniquely first-person act: the ability to test whether *one's own* rule of action could be willed as a universal law. This is an act of autonomous, *self-conscious* agency that a purely rule-following system cannot perform, since this act presupposes the capacity to distinguish one's own *subjective* maxim from an objective

[11] Several scholars, many of whom are contributors to the same edited volume, already pointed to AI's inability to self-ascribe its output as the primary ground for denying it genuine agency. See: S. Baiasu, *The Challenge of (Self-)Consciousness: Kant, Artificial Intelligence and Sense-Making*, in H. Kim, D. Schönecker (edited by), *Kant and Artificial Intelligence*, Walter de Gruyter GmbH, Berlin-Boston 2022, p. 125; L. Benossi S. Bernecker, *A Kantian Perspective on Robot Ethics*, in H. Kim, D. Schönecker (edited by), *Kant and Artificial Intelligence*, Walter de Gruyter GmbH, Berlin-Boston 2022, p. 157; R. Evans, *The Apperception Engine*, in H. Kim, D. Schönecker (edited by), *Kant and Artificial Intelligence*, cit., p.85; Walter de Gruyter GmbH, Berlin-Boston 2022, p. 157; D. Schönecker, *Kant's Argument from Moral Feelings: Why Practical Reason Cannot Be Artificial*, in H. Kim, D. Schönecker (edited by), *Kant and Artificial Intelligence*, Walter de Gruyter GmbH, Berlin-Boston 2022, pp. 180-181.

moral law.

It also follows that an AI lacking self-consciousness cannot be recognized as a bearer of legal rights. This is because, Kant's *Universal Principle of Right*, the very ground of such rights, applies only to actions resulting from "the free use" of choice. The principle states that an action is right if it can coexist with «the freedom of everyone» (MM; VI:231). An entity without self-consciousness cannot be a subject of rights, since it is not an autonomous agent capable of free choice. For something to be an object of free choice, a subject must first be able to posit it as an end for herself. This capacity to set ends, however belongs to the faculty of self-consciousness and is thus fundamentally distinct from the mere combination of sense perceptions[12].

4. Conclusion

The most important findings of this article can be summarized in the following two theses:

A. If an AI is capable of identifying objects in images by applying concepts, then it satisfies the Kantian condition for consciousness.

B. Acknowledging a being as conscious does not, however, entail granting of moral and legal rights to it.

[12] This Kantian analysis, of course, is not intended to exclude other philosophical approaches to the ethical status of AI. Indeed, there is a growing body of research dedicated to evaluating AI's potential as an artificial moral agent (AMA). For some key contributions to this discussion, see: K. E. Himma, *Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?*, in «Ethics and Information Technology», n. 11, 2009, pp. 19-29; M. Hunyadi, *Artificial Moral Agents. Really?*, in Jean-Paul Laumond, Emanuelle Danblon, Céline Pieters (edited by), *Wording Robotics: Discourses and Representations on Robotics*, Springer, Toulouse 2019, 59-69; A. Martinho, A. Poulsen, M. Kroesen, C. Chorus, *Perspectives about Artificial Moral Agents*, in «AI and Ethics», n. 1, 2021, pp. 477-490; M. Arvan, *Varieties of Artificial Moral Agency and the New Control Problem*, in « Humana.Mente - Journal of Philosophical Studies», n. 15, 2022, pp. 225-256; S. Nyholm, *Is Academic Enhancement Possible by Means of Generative AI-Based Digital Twins?*, in «The American Journal of Bioethics», n. 23, 2023, pp. 44-47; J. Graff, *Moral Sensitivity and the Limits of Artificial Moral Agents*, in «Ethics and Information Technology», n. 26, 2024, pp. 1-13; R. van Woudenberg, C. Ranalli, D. Bracker, *Authorship and ChatGPT: a Conservative View*, in «Philosophy & Technology», n. 37, 2024, pp. 1-26; J. McLoughlin, *The Work of Art in the Age of Artificial Intelligibility*, in «AI & Society», n. 40, 2025, pp. 371-383.

The policy convergence of restrictive abortion laws and autonomous vehicle regulation in the U.S.[a]

Siobhain Lash[*]

Abstract

In questo articolo sostengo che le discussioni sulle normative relative alle auto a guida autonoma e sulle attuali leggi restrittive in materia di aborto negli Stati Uniti siano intersezionali. Queste discussioni hanno gravi implicazioni per l'autonomia fisica, la libertà di movimento e la sorveglianza delle donne e delle comunità emarginate e minoritarie. In tutto l'articolo, concentro la mia discussione sull'intersezione tra leggi restrittive in materia di aborto e il divieto di guidare per gli esseri umani. Esamino la possibilità di vietare i conducenti umani e come potrebbe essere un futuro senza conducenti secondo Sparrow e Howard, nonché la transizione verso tale futuro. Successivamente, metto in evidenza i quadri tecnologici e politici che potrebbero influenzare la restrizione del diritto fondamentale di viaggiare e i precedenti costituzionali e giuridici. L'obiettivo del mio articolo è mostrare come le discussioni sulle normative relative alle auto a guida autonoma e le attuali leggi restrittive sull'aborto negli Stati Uniti siano intersecanti e sottolineare le conseguenti gravi implicazioni politiche.

Parole Chiave: regolamentazione dell'IA, veicoli a guida autonoma, sorveglianza digitale, autonomia del corpo.

In this paper, I argue that discussions of self-driving car regulations and current restrictive abortion laws across the United States are intersectional. These discussions have serious implications for bodily autonomy, freedom of mobility, and surveillance for women and marginalized and minority communities. Throughout the paper, I center my discussion on the intersection of restrictive abortion laws and the banning of human drivers. I examine the possibility of banning human drivers and what a driverless future looks like according to Sparrow and Howard, and the transition to such a future. Then, I highlight the technological and policy frameworks that could inform restricting the fundamental right to travel and the constitutional and legal precedents. The goal of my paper is to show how discussions of self-driving car regulations and current restrictive abortion laws across the United States intersect and to emphasize the subsequent serious policy implications.

Keywords: AI regulation, autonomous vehicles, digital surveillance, bodily autonomy.

[a] Received on 19/02/2025 and published on 09/12/2025.
[*] West Virginia University, e-mail: siobhain.lash@mail.wvu.edu.

1. Introduction

In this paper, I argue that discussions of self-driving car regulations and current restrictive abortion laws across the United States converge in unexpected but important ways. In particular, these discussions have serious implications for bodily autonomy, freedom of mobility, and digital surveillance for women and marginalized and minority communities. Current ethical debates around self-driving cars involve questions related to what, if any, type of ethical settings self-driving cars ought to possess, like personal ethical settings (PES) or mandatory ethical settings (MES)[1].

Other discussions include the potential banning of human drivers altogether. In their article, "When Human Beings are Like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport," Robert Sparrow and Mark Howard[2] advocate for mandatory self-driving cars, which effectively outlaws manually-driven cars. They cite the increase in both pedestrian and driver fatalities within the past ten years and to the likelihood that humans will jettison road safety laws and regulations as production and use of self-driving cars grow. The ethical debates surrounding driverless vehicles contain salient policy implications, especially when considered in combination with current laws around digital surveillance and bodily autonomy.

Consequently, for this paper, I examine the real-world policy implications of Sparrow and Howard's argument when understood as happening within broader policy contexts, such as the rise in abortion bans across the U.S., including restrictions on interstate travel, and the increase in digital surveillance. While developers have not fully realized autonomous vehicles yet, these rapidly progressing technologies are already reshaping future discussions on transportation, technology, and surveillance. Thus, I specifically examine Sparrow and Howard's argument in relation to restrictive abortion laws across the United States since the overturning of Roe v Wade in June 2022. Ultimately, I reject Sparrow and Howard's proposal, and others like it, on normative grounds. In particular, I argue that their vision for the banning of human drivers relies too heavily on forms of digital surveillance and control over mobility that is morally impermissible. I claim that these practices violate core ethical principles like consent and also establish dangerous precedents that corporations, law enforcement agencies, and political actors could exploit to target vulnerable populations in the future.

For this paper, I am assuming that these laws are overly restrictive. Additionally, I take a similar approach to policy that Gerald Gaus takes in *The Tyranny of the Ideal: Justice in a Diverse Society.* In *Tyranny of the Ideal*[3], Gaus acknowledges that policy frameworks in complex societies can significantly inform other policies and overlap in unexpected ways. This is a stance this paper takes.

Section 2 discusses current policy proposals for regulating driverless cars and the transition to a total ban on human drivers. Section 3 provides an analysis of current policies

[1] J. Gogoll, J.F. Müller, *Autonomous Cars: In Favor of a Mandatory Ethics Setting*, in «Science and Engineering Ethics», XXIII, n. 3, 2017, pp. 681-700.
[2] R.J. Sparrow, M. Howard, *When Human Beings Are like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport*, in «Transportation Research Part C: Emerging Technologies», LXXX, 2017, pp. 206-215.
[3] G. Gaus, *The Tyranny of the Ideal: Justice in a Diverse Society*, Princeton University Press, Princeton 2016.

related to bodily autonomy and interstate mobility post Roe v Wade. Section 4 examines the convergence of post-Roe abortion bans with constitutional law, mobility rights, and state authority in the United States, with a focus on how states might leverage legal precedents and existing surveillance technologies to restrict or regulate interstate travel for abortion services. Section 5 highlights how data collection and the current digital infrastructure might be weaponized to monitor movement and enforce future mobility restrictions and the regulation of autonomous vehicles. Section 6 explores the practical and ethical implications of the overlap between proposed and current abortion laws and evolving regulations of autonomous drivers, including a case study to illustrate the potential consequences of these regulatory frameworks meeting. Section 7 concludes the paper.

2. Regulating Driverless Cars

There are many ethical reasons to support outlawing human drivers and promoting driverless cars and autonomous vehicles (AV)[4][5] beyond the fatal crash statistics[6]. However, it remains a complex policy issue that includes questions of implementation and what the transition entails. In their article, «When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport» Robert Sparrow and Mark Howard discuss the ethical implications of driverless vehicles and their implementation. Sparrow and Howard argue that the moment that driverless cars are safer than human drivers, that human drivers ought to be outlawed. «[W]e argue that the invention of fully autonomous vehicles that pose a lower risk to third parties than human drivers will establish a compelling case against the moral permissibility of manual driving»[7]. They bolster their argument through observing that morally allowing a human to use an AV with at least level 3 automation puts the rest of the drivers at unnecessary risk. Because the AV will be reliably autonomous, the person may mistakenly rely on its ability to encounter all and any risks. Subsequently, this may encourage the human driver to engage in reckless behavior, such as reading, drinking, having sex, falling asleep while at the wheel, or parents sending kids alone to school[8].
Additionally, the independence that such AVs provide could encourage individuals previously deemed unfit to drive to feel embolden to get behind the wheel, such as those

[4] There are five levels, which range from levels 0-5. Level 0 denotes a vehicle that does not contain any sort of automation. Levels 1-3 have some level of the AI performing dynamic driving tasks (DDT). DDT is when an automated system partially or conditionally performs operational and tactical functions like steering, acceleration, and braking (Y. Freemark *et al.*, *Regulations to Respond to the Potential Benefits and Perils of Self-Driving Cars*, Metropolitan House and Communities Policy Center, September 2022, pp. 5-6). Levels 4-5 entail an automated system fully or highly performing all DDTs under relevant or all conditions (*ivi*, p. 6). The vehicle moves from an AV to a driverless vehicle when its automated system fully performs DDTs under all conditions and does not need any human intervention or external override for strategic functions.
[5] It is important to note that car manufacturers, like GM, are currently planning for AVs that do not have steering wheels (G. Barta, *GM CEO Mary Barra Doubles Down On AV With No Steering Wheel*, in «GM Authority», 31 October 2024: https://gmauthority.com/blog/2024/10/gm-ceo-mary-barra-doubles-down-on-autonomous-vehicle-with-no-steering-wheel/). So, a human overriding the system to drive physically will not be an option.
[6] J. Gogoll, J.F. Müller, *Autonomous Cars: In Favor of a Mandatory Ethics Setting*, cit., p. 682.
[7] R.J. Sparrow, M. Howard, *When Human Beings Are like Drunk Robots*, cit., p. 207.
[8] Ivi, p. 208.

with cognitive impairments or other medical conditions that compromise the individual's ability to safely drive.

Sparrow and Howard argue that this would increase the number of unfit drivers on the road. They note[9], «Even if the vehicle can provide several seconds of warning before requiring manual control to be re-engaged, then, there may be no one in a fit state to do so» Sparrow and Howard continue to argue that such a transitional design[10] would render the human driver as morally equivalent to a drunk robot. They claim, «Moreover, imposing this extra risk on third parties will be unethical: the human driver will be the moral equivalent of a drunk robot. Eventually, we believe, the compelling moral argument against human drivers will be reflected in law: driving will be made illegal»[11]. Sparrow and Howard argue that a transitional AV would increase both reckless and unfit drivers, making most drivers morally equivalent to drunk robots. They argue that this comparison would provide robust support for the inevitable outlawing of human drivers, rendering driving illegal.

Sparrow and Howard's discussion highlights a difficulty that may arise during the transition from AVs to driverless vehicles. Currently, there are a handful of empirical examples of Tesla drivers that display the type of reckless behavior that is the focus of Sparrow and Howard's discussion. For example, in February of 2024, a video went viral of a Tesla driver in Full-Self Driving (FSD) mode wearing Apple's new virtual reality (V.R.) headset. Although the video went viral, the trend was not widespread, it still prompted United States Secretary of Transportation Pete Buttigieg and the National Highway Traffic Safety Administration to respond to the social media posts[12]. Similarly, in April 2024, a Tesla driver in FSD mode hit and killed a motorcyclist due to being on his phone at the time of the collision. Due to being distracted, the driver was not able to manually respond and override the system to prevent the collision[13]. Despite these instances of reckless driving Sparrow and Howard warn against, they may not reflect future driver behavior once AVs do become more widely used. However, they do not inspire support for retaining human driving, either. Even so, Howard and Sparrow's discussion does highlight important potential public health and safety concerns related to the possible increase in drunk robots on the road with the increased availability of AVs.

The upshot is that although autonomous vehicle drivers still only marginally make up a small fraction of drivers in the US, the push for more commercially available AVs

---

[9] *Ibidem*.

[10] Although Sparrow and Howard do not explicitly call it a transitional design, their discussion is in relation to a transitional phase that is between fully autonomous vehicles and those that still require some level of human intervention. So, for clarity, I use transitional design to denote this transitional period.

[11] Ivi, p. 209.

[12] J. Jiménez, *Stop Wearing Vision Pro Goggles While Driving Your Tesla, U.S. Says*, in «The New York Times», 6 February 2024: https://www.nytimes.com/2024/02/06/technology/personaltech/apple-vision-pro-tesla.html; D. Shepardson, *Viral Videos of Tesla Drivers Using VR Headsets Prompt US Government Alarm*, in «Reuters», 5 February 2024: https://www.reuters.com/business/autos-transportation/us-transport-chief-urges-drivers-pay-attention-after-virtual-reality-driver-2024-02-05/.

[13] H. Jin, *Tesla Car That Killed Seattle Motorcyclist Was in 'Full Self-Driving' Mode, Police Say*, in «Reuters», 31 July 2024: https://www.reuters.com/business/autos-transportation/tesla-was-full-self-driving-mode-when-it-hit-killed-seattle-motorcyclist-police-2024-07-31/; M. Lenthang, *Tesla in Fatal Seattle-Area Crash That Killed Motorcyclist Was in Self-Driving Mode, Officials Say*, in «NBC News», 31 July 2024: https://www.nbcnews.com/news/us-news/tesla-fatal-seattle-area-crash-killed-motorcyclist-was-self-driving-mo-rcna164488.

increases the risk of drunk robots on the road. This is because, as car manufacturers work towards fully driverless vehicles, the existing AVs will be transitional models. Sparrow and Howard warn that even though the driver still has to intervene and take manual control at a moment's notice, similar to traditional vehicles, these transitional models could entice drivers to engage in reckless behavior. Consequently, the transition from AVs to driverless vehicles could potentially encourage policymakers to use Sparrow and Howard's argument to inform their case against the moral permissibility of manual driving. As a result, by the time driverless vehicles are commercially available, then regulations may already be in place that fully ban human drivers.

3. Current Policies: Bodily Autonomy, Surveillance, and Interstate Mobility

a. The Overturning of Roe v Wade 2022

The questions around bodily autonomy and freedom of mobility with the advent of driverless vehicles could find precedence or traction in the overturning of Roe v Wade, otherwise known as the Dobbs decision. In June 2022, the United States Supreme Court Justices overturned Roe v Wade, initiating the automatic banning of abortion in trigger law states.[14] Since then, lawmakers have introduced and passed an onslaught of various anti-abortion laws and ordinances, further restricting access to abortion and abortion-adjacent medical care to millions of people of reproductive age. As a result, 14 states have total or near-total abortion bans and a total of 41 states have some sort of abortion ban[15]. Only 9 states do not have any gestational limits or other restrictions to access an abortion[16]. In some states, anti-abortion and far-right organizations have continued their efforts to ban most reproductive care through introducing legislation that attempts to ban the abortion pill, plan b, birth control, and telemedicine in the context of offering abortion care[17]. In

[14]These states include Arkansas, Idaho, Kentucky, Louisiana, Mississippi, Missouri, North Dakota, Oklahoma, South Dakota, Tennessee, Texas, Utah, and Wyoming (J. Jiménez, *What Is a Trigger Law? And Which States Have Them?*, in «The New York Times», 4 May 2022: https://www.nytimes.com/2022/05/04/us/abortion-trigger-laws.html; Center for Reproductive Rights, *Abortion Laws by State*, 2024: https://reproductiverights.org/maps/abortion-laws-by-state/; E. Nash, I. Guarnieri, *13 States Have Abortion Trigger Bans – Here's What Happens When Roe Is Overturned*, Guttmacher Institute, 6 June 2022: https://www.guttmacher.org/article/2022/06/13-states-have-abortion-trigger-bans-heres-what-happens-when-roe-overturned).

[15] Guttmacher Institute, *State Bans on Abortion Throughout Pregnancy*, 29 July 2024: https://www.guttmacher.org/state-policy/explore/state-policies-abortion-bans.

[16] *Ibidem.*

[17] M. Kekatos, *A State-by-State Breakdown of Abortion Laws 2 Years after Roe Was Overturned*, in «ABC News», 22 June 2024: https://abcnews.go.com/US/state-state-breakdown-abortion-laws-2-years-after/story?id=111312220; J. Biden, *Statement from President Joe Biden on Senate Republicans Blocking Efforts to Safeguard Nationwide Access to Contraception*, The White House, 5 June 2024; A. Nawaz, S. Khan, *Louisiana Restricts Access to Abortion Pills by Classifying Them as a Controlled Substance*, in «PBS News», 24 May 2024: https://www.pbs.org/newshour/show/louisiana-restricts-access-to-abortion-pills-by-classifying-them-as-a-controlled-substance.

addition, states such as Alabama, Texas, Idaho, Tennessee, and Oklahoma have proposed or passed bills restricting women's interstate mobility[18].

To illustrate these legislative actions, consider conservative stronghold Amarillo, Texas. Amarillo became one of the first cities to attempt to ban interstate travel. In July 2024, the people of Amarillo voted to put on the November ballot Proposition A, an ordinance that would outlaw people from using local streets and highways as a way to obtain through an abortion in New Mexico, where abortion is still legal[19]. On November 5th, 2024, Amarillo residents resoundingly and unexpectedly rejected Proposition A. Because the questionable constitutionality of these laws, even conservative strongholds like Amarillo are resisting attempts for anti-abortion organizations to propose or implement such draconian measures. Given this, anti-abortion groups are having to pursue creative ways to circumvent any legal barriers and public backlash. The strategy is to use duplicitous language to make the laws more palatable to voters and constituents, and thus to circumvent the resistance they are currently receiving.

For example, some lawmakers are including abortion travel bans under trafficking laws under the pretense to safeguard minors. These laws have been called "abortion trafficking laws"[20]. Anti-abortion organizations are naming similar ordinances, like the Amarillo ordinance up for vote in November 2024, as "sanctuary city for the unborn" ordinances. Other laws include bans on "abortion tourism,"[21] or "Bans Offshore Abortion Tourism Act" also referenced to by its acronym BOAT[22]. BOATs, trafficking laws, and sanctuary ordinances are popping up across the country that fall under the radar due to their misleading names and goals[23]. For example, a constituent does not need to be "pro-life" to find it difficult to justify voting against a law that presumably aims to address child sex trafficking in their area, especially if they are a parent.

Given this, if lawmakers fail, either intentionally or not, to fully inform constituents about the law's actual intent, the inclusion of abortion travel bans in trafficking legislation could succeed. Such legislation can serve as the legal foothold that anti-abortion organizations need to advance these policies. This insistence on passing abortion travel bans, despite probable public backlash, reflects a broader legal trend that Justice Kavanaugh, for example, acknowledged in his decision to vote to overturn Roe v Wade. Kavanaugh noted that banning interstate travel for abortion was a likely next step.

---

[18] M.A. Pazanowski, *Doubts Over Abortion Travel Bans Lead States to Try Other Means*, in «Bloomberg Law», 15 May 2024: https://news.bloomberglaw.com/health-law-and-business/doubts-over-abortion-travel-bans-lead-states-to-try-other-means.

[19] C. Sherman, *Texas City to Vote on Ban on People Helping Patients Traveling for Abortion*, in «The Guardian», 19 July 2024: https://www.theguardian.com/us-news/article/2024/jul/19/texas-abortion-travel-ban; J.L. Carver, *Amarillo City Council Must Vote on Abortion Travel Ban Following Successful Voter Petition*, in «The Texas Tribune», 16 May 2024: https://www.texastribune.org/2024/05/16/amarillo-texas-abortion-travel-ban-vote/.

[20] Idaho State Legislature, *Idaho Code § 18-623*, 2024; B. Pierson, *Idaho Seeks to Revive 'abortion Trafficking' Law in US Appeals Court*, in «Reuters», 7 May 2024: https://legislature.idaho.gov/statutesrules/idstat/title18/t18ch6/sect18-623/.

[21] Abortion tourism is a term used to denote when a person travels outside of a state with draconian abortion laws to seek an abortion or abortion-related medical care in a state where abortion is legal.

[22] R.K. Weber, *H.R. 5319 - Ban Offshore Abortion Tourism Act*, 118th Congress, 29 August 2023: https://www.congress.gov/bill/118th-congress/house-bill/5319.

[23] ACLU of Illinois, *Sanctuary for the Unborn Ordinances*, 12 October 2023: https://www.aclu-il.org/en/campaigns/sanctuary-unborn-ordinances.

However, he remained confident that such bans would face significant criticism and ruled as unconstitutional. So far, Kavanaugh has been right. Nevertheless, such a ban has precedence and a policy infrastructure from which contemporary proponents of such bans can build their current policies. While it can seem impossible for such blatantly unconstitutional laws or ordinances to pass, these organizations are adept at exploiting legal frameworks to advance or legitimize them. These efforts become significantly reinforced when endorsed by institutions like the Supreme Court or other federal entities.

Additionally, these organizations could exploit state legal frameworks, such as presumptive extraterritorial power that states have. Given this, states have some power to enforce regulations against a citizen, even if that citizen is in a different state. This is challenging for many reasons, some obvious, but the bottom-line is that states do have presumptive power to regulate their citizen's movements and behaviors, even when that citizen either relocates or travels to a different state. The enforceability of the regulation depends on how the state is interpreting both the law and the individual's behavior, but there have been some instances where an individual was subject to the regulations of state A, despite being in state B. Again, states do not usually exercise this power given the backlash and resistance they are likely to receive from the public.

4. Abortion Bans and Travel Restrictions

The tension between individual interstate mobility and state jurisdiction has become even more pronounced after the Dobbs decisions. Proposals for abortion travel bans have led experts and journalists to explore further downstream implications of the decision, such as possible implementation and enforceability of abortion travel bans. Some experts[24] focused on the presumed fundamental right US citizens have regarding their freedom of interstate mobility or travel, as possible constitutional and legal precedents for anti-abortion advocates. In their congressional report, "Congressional Authority to Regulate Abortion," attorneys Kevin J. Hickey and Whitney K. Novak[25] note that anti-abortion advocates have previously used the Commerce Clause, in particular, to support abortion-related legislation. They note that states could similarly use the clause to extend to abortion services, interpreting them as commercial activities that engage in interstate commerce[26].

Additionally, other discussions include other legal foundations, like the Model Penal Code[27] and the Spending clause to restrict interstate travel for abortion services[28]. Americans assume that the Fourteenth Amendment Article IV and the Privileges and Immunities Clause of Article IV protects the right to travel, but this is not the case. The Fourteenth Amendment does not mention the explicit right to travel. Rather, prevailing interpretations have presumed that interstate mobility is concomitant to the other rights guaranteed in the Fourteenth Amendment. Furthermore, some interpretations of the Right

[24] D.K. Brown, *Extraterritorial State Criminal Law, Post- DOBBS*, University of Virginia School of Law, 28 August 2023: https://www.law.virginia.edu/node/2171216; K.J. Hickey, W.K. Novak, *Congressional Authority to Regulate Abortion*, CRS Report LSB10787, 2022.
[25] K.J. Hickey, W.K. Novak, *Congressional Authority to Regulate Abortion*, cit.
[26] Ivi, p. 3.
[27] D.K. Brown, *Extraterritorial State Criminal Law, Post- DOBBS*, cit.
[28] I. Millhiser, *The Unconstitutional Plan to Stop Women from Traveling out of State for an Abortion, Explained*, in «Vox», 12 September 2023: https://www.vox.com/23868962/texas-abortion-travel-ban-unconstitutional.

to Travel and Privileges and Immunities Clause in Article IV of the Constitution include believing it grants and guarantees a US state citizen the immunities and privileges of citizens in several states[29]. However, the Constitution does not enshrine the fundamental right to travel. Rather, an overlapping of different federal and state policies and laws have led to *interpretations* by states and the U.S. Supreme Court. The U.S. Supreme court, for example, recognizes the freedom of travel as a fundamental right.

Nevertheless, the federal and state governments can restrict interstate mobility due to public health and national security reasons[30]. Even so, the prevailing view is that despite these possible constitutional and legal precedents, such travel bans would receive extensive challenges, and have so far. Despite these challenges, I briefly cover other legal precedents, specifically around surveillance that nefarious actors could use. There are potentially two types of surveillance that anti-abortion groups could pursue to advance their goals. They could either use mobility monitoring, which does have a legal precedence in the United States or digital surveillance, which is a growing industry.

a. Monitoring Mobility

Currently, the United States uses various tools and systems of monitoring mobility. The most well-known form is Location Monitoring (LM), which involves the use of electronic monitors that use Global Positioning Systems (GPS). Typically, an LM device attaches to an individual's ankle for 24/7 surveillance. These devices notify officers of when the individual with the LM device moves outside of the set parameters or if the individual tampers with the device. Similarly, it monitors whenever the individual leaves or enters the designated destination[31]. Other LM devices include data or cellular signals, voice recognition (VR), and virtual monitoring, like tracking the individual through apps on their phone. There are several LM restriction levels. They include, curfew, home detention, home incarceration, and stand-alone monitoring[32].

Additionally, the US restricts the mobility of formerly incarcerated individuals known as probation. The release of formerly incarcerated individuals comes with the expectation of receiving community supervision. Under some conditions of probation, individuals may be asked to «remain within the jurisdiction of the court, unless granted permission to leave by the court or a probation officer»[33]. Ultimately, for the individual under probation, the courts decide where they can and cannot travel and what restrictions to enforce. Consequently, despite the complexity and challenges of enforcing abortion travel bans, anti-abortion advocates can draw on constitutional and legal precedents related to monitoring mobility. Correspondingly, the criminal justice system contains established protocols and procedures that require restricting an individual's mobility, which future policies can expand upon.

---

[29] *Right to Travel and Privileges and Immunities Clause*, in *Constitution of the United States of America: Analysis and Interpretation*, 2024.
[30] *Interstate Travel as a Fundamental Right*, in *Constitution of the United States of America: Analysis and Interpretation*, 2024; Legal Information Institute, *Interstate Travel*, Cornell Law School, 2024.
[31] United States Courts, *How Location Monitoring Works*, March 2023: https://www.uscourts.gov/services-forms/probation-and-pretrial-services/supervision/how-location-monitoring-works.
[32] *Ibidem.*
[33] United States Courts, *Chapter 3: Location Monitoring (Probation and Supervised Release Conditions)*, 2024.

b. Digital Borders

In tandem with traditional physical mobility monitoring, the U.S. has increasingly turned to the use of digital borders as a means to control movement. The evolving technologies and biometrics of digital borders require a flexible definition that captures their intent and purpose. In their article, "The digital border: Mobility beyond territorial and symbolic divides" Lilie Chouliaraki and Myria Georgiou[34] offer a clear definition of digital borders and state:

> The digital border, as we have established, can be grasped as a shifting assemblage of technologies and meanings organised around historically-specific power relations that regulate migrant mobility across the binary of inside/outside *at* the edge and *within* the boundaries of national sovereignty.

Digital borders use technology to regulate migrant mobility across borders or within a country. This trend is not isolated to the U.S., as the European Union (EU) is in the process of implementing iCROSS, which is a digital and portable border control system[35] and has other digital processes in place to monitor and track travel across the continent.

Even so, after taking office in 2025, the Trump administration immediately initiated a "social media screening/ vetting" policy for travelers, specifically visa or green card holders, to show their social media to Customs and Border Patrol (CBP). While this is not a new policy, it is a more subjective and invasive form of the Electronic System for Travel Authorization (ESTA), developed by the Department of Homeland Security (DHS) in 2008[36]. These enforcement changes have already had considerable consequences. Several high-profile cases have involved foreign nationals denied entry to the United States while attempting to attend conferences or go on holiday. According to multiple reports, this was after criticizing or mocking either Trump or his administration. This has increased significantly with CBP recently requiring F, M, and J visa applicants to have all of their social media platforms public or risk rejection[37]. In addition, on June 26th, 2025, the Trump administration announced that green card holders could have their eligibility to remain in the U.S. revoked if they support terrorism or violence[38]. According to several reports, the Trump administration has so far interpreted "support of terrorism or violence" as support for Palestine or criticizing Israel[39].

---

[34] L. Chouliaraki, M. Georgiou, *The Digital Border: Mobility beyond Territorial and Symbolic Divides*, in «European Journal of Communication», XXXIV, n. 6, 2019, pp. 594-605: 600.

[35] European Commission, *Intelligent Portable Border Control System*, CORDIS, 6 September 2024: https://cordis.europa.eu/project/id/700626.

[36] U.S. Customs and Border Protection, *Strengthening Security of the VWP through Enhancements to ESTA*, 11 February 2025: https://www.cbp.gov/travel/international-visitors/esta/enhancements-to-esta-faqs.

[37] U.S. Embassy in Mali, *Updated Social Media Disclosure Requirement for F, M, J Visa Applicants*, 24 June 2025; United States Department of State, *Announcement of Expanded Screening and Vetting for Visa Applicants*, 18 June 2025: https://ml.usembassy.gov/u-s-requires-public-social-media-settings-for-f-m-and-j-visa-applicants/.

[38] B. Rahman, *New Warning Issued to Green Card Holders*, in «Newsweek», 26 June 2025: https://www.newsweek.com/green-card-warning-issued-immigration-violence-terrorism-2091042.

[39] A. Mahdawi, *Keep Calm (but Delete Your Nudes): The New Rules for Travelling to and from Trump's America*, in «The Guardian», 15 May 2025: https://www.theguardian.com/us-news/2025/may/15/travel-trump-

For example, in March 2025, immigration officers denied a French scientist heading to a conference in Houston over his criticism of the Trump administration[40]. In June 2025, CBP went through an Australian writer's phone and used what they found on it as a pretense to detain him in Los Angeles. He was then deported back to Melbourne because of his coverage of the Pro-Palestinian protests in 2024 while he was a student at Columbia[41]. These cases reflect a broader trend among a growing number of unreported detentions and deportations by the CBP at the border.

In light of these high-profile cases, the Trump administration's enforcement and interpretation of the ESTA policy has drawn renewed public scrutiny, particularly over concerns of the politicized use of digital surveillance to target political dissidents rather than genuine national security threats. These concerns intensified in May 2025, after the Trump administration announced they formed a partnership with surveillance and technology firm Palantir to compile profiles on every American[42]. Palantir also revealed it is developing a platform to track migrant movements in real time for Immigration and Customs Enforcement (I.C.E.)[43]. Alongside creating these databases on Americans, the Trump administration also adopted Foundry, a technology program that specializes in organizing and analyzing data to facilitate information sharing between different agencies[44]. In addition, in June 2025, the Trump administration created a new US army division, Detachment 201, which is the Army's Executive Innovation Corps, designed to "fuse cutting-edge expertise with military innovation." According to the Army's official website, the division's new Army Reserve Lt. Cols. include, «Shyam Sankar, Chief Technology Officer for Palantir; Andrew Bosworth, Chief Technology Officer of Meta; Kevin Weil, Chief Product Officer of OpenAI; and Bob McGrew, advisor at Thinking Machines Lab and former Chief Research Officer for OpenAI»[45]. Given the Trump administration's approach to surveillance and dissent, the formation of the new tech army division further

---

america-us-border-detentions; A. Greenberg, M. Burgess, *How to Enter the US With Your Digital Privacy Intact*, in «Wired», 24 March 2025: https://www.wired.com/2017/02/guide-getting-past-customs-digital-privacy-intact/; M. Malaver, J. Weaver, *US Border Officials Can Search Your Phone without a Warrant. What to Know*, in «Tampa Bay Times», 20 April 2025: https://www.tampabay.com/news/2025/04/20/immigration-phone-searches-customs-border-protection-florida/.

[40] R. Mackey, *French Scientist Denied US Entry after Phone Messages Critical of Trump Found*, in «The Guardian», 19 March 2025; Reuters, *French Scientist Denied Entry into the US, French Government Says*, 20 March 2025: https://www.theguardian.com/us-news/2025/mar/19/trump-musk-french-scientist-detained.

[41] A. Lewis, *Australian Denied Entry to US after Being Grilled on Israel-Gaza Views*, in «ABC News», 15 June 2025: https://www.abc.net.au/news/2025-06-16/australian-denied-entry-united-states-israel-gaza-columbia/105419154.

[42] S. Frenkel, *Lawmakers Demand Palantir Provide Information About U.S. Contracts*, in «The New York Times», 17 June 2025: https://www.nytimes.com/2025/06/17/technology/palantir-government-contracts-democrats-letter.html; J. Jones, *Trump Appears to Be Building an Unprecedented Spy Machine That Could Track Americans*, in «MSNBC», 30 May 2025; B. Allyn, *How Palantir, the Secretive Tech Company, Is Rising in the Trump Era*, in «NPR», 3 May 2025: https://www.npr.org/2025/05/01/nx-s1-5372776/palantir-tech-contracts-trump.

[43] S. Frenkel, A. Krolik, *Trump Taps Palantir to Compile Data on Americans*, in «The New York Times», 30 May 2025: https://www.nytimes.com/2025/05/30/technology/trump-palantir-data-americans.html.

[44] *Ibidem.*

[45] U.S. Army Public Affairs, *Army Launches Detachment 201: Executive Innovation Corps to Drive Tech Transformation*, 13 June 2025: https://www.army.mil/article/286317/army_launches_detachment_201_executive_innovation_corps_to_drive_tech_transformation.

raises concerns about the lack of oversight and the scale and scope of digital surveillance. These concerns extend not only to targeting alleged security threats, but also to Americans or others seemingly targeted by the administration.

5. Emerging Telematics Systems Technology in Automobiles

In this section, I argue that widespread, unregulated surveillance built into emerging telematics technologies poses serious ethical risks for women and other historically marginalized communities. I draw on digital ethics frameworks to emphasize ethical concerns regarding autonomy, disproportionate impact on populations already marginalized and heavily surveilled by current institutions, and privacy violations within the larger context of reproductive rights. From this perspective, I show how these surveillance technologies coupled with the increase in mobility monitoring occur due to morally impermissible means. This is because data collection and the data broker industry are both heavily unregulated. This has allowed companies from Meta to car manufacturers to engage in nontransparent practices that include the implicit collection of data of their consumers.

Recently, car manufacturers across the board have embedded telematics systems or software over-the-air (SOTA) technology into newer models of their cars. According to Rambus Press[46], over-the-air programming refers, "to the ability to download applications, services, and configurations over a mobile or cellular network. Over-the-air (OTA) programming is used to automatically update firmware, software, and even encryption keys." OTA technologies allow car manufacturers to connect directly to the cars, either to disable certain functions in the car to prevent that tethered feature unless the individual pays the required subscription fee or to update certain upgrades. Tesla was the first automaker to monetize OTAs in early 2019. Other luxury brands, including Mercedes, BMW, Audi, and Lexus have all rolled out their subscription services.

For example the new A3 requires users to choose a subscription package to access, «adaptive cruise control, Apple CarPlay and Android Auto, automatic high beams, and, bafflingly, dual-zone climate controls»[47]. Lexus rolled out its "The Go Anywhere Plan" that provides consumers access to features originally part of the car, such as navigational systems, remote access to the car, like un/locking the doors, and safety features within the car. If the consumer does not purchase the premium package, then certain safety features, along with the sat nav, will be disabled[48]. BMW requires an $18-a-month subscription for consumers to access their heated seats and remote-start features, features their cars already contain[49]. This is just one of the potential (mis)use of systems that monitor mobility in cars to collect data for training and other purposes specific to the manufacturer.

[46] Rambus, *What Is OTA in Automotive? Over the Air Updates Explained*, 13 May 2022: https://www.rambus.com/blogs/ota-updates-explained/.
[47] C. Teague, *The New Audi A3 Comes with Subscription Fees in Europe*, in «The Truth About Cars», 12 March 2024: https://www.thetruthaboutcars.com/cars/news-blog/the-new-audi-a3-comes-with-subscription-fees-in-europe-44505658.
[48] Lexus, *Subscription Plans*, 23 January 2024: https://support.lexus.com/s/article/Subscription-Plans-L.
[49] H.D. Beaver, *Automakers' Added Subscription Fees Raise Legal Questions*, in «Kiplinger», 2 January 2024: https://www.kiplinger.com/personal-finance/automakers-added-subscription-fees-raise-legal-questions.

a. *Digital Surveillance*

For cars to continue advancing toward full autonomy, as envisioned by Howard and Sparrow, they require collecting a vast amount of consumer data which often happens without the consumer's knowledge or consent, and commonly car manufacturers are not transparent about how, why, when, or what kind of data they collect. This is in addition to the amount of data consumers already have collected on and about them throughout the day, including on social media, their medical apps, like MyChart, the weather app, security cameras, and voice assistants[50]. This is known as datafication. In their article, "How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches" Oskar J. Gstrein and Anne Beaulieu explore the increase in datafication of our private and public spheres. Gstrein and Beaulieu state:

> Datafication is a complex process that is often associated with digital technologies and (Big) data infrastructure. Furthermore, data and data-related capabilities are central to datafication. This includes the extension of automation, the proliferation of digital technologies, the willing production of massive amounts of data and the combination and circulation of datasets[51].

Datafication requires an interconnectedness between digital devices, supported by both digital infrastructures and societal practices. Thus far, I have shown multiple ways that companies and the government in the United States collect data on the public, from ankle monitors to social media posts and the use of surveillance technologies. Datafication allows companies like Palantir to quantify, track, and render things like social ties and associations into digital data and traces. Gstrein and Beaulieu state:

> Hence, data generated by highly connected and monitored environments do not only allow one to make explicit statements about the individuals consenting to its collection and use, but also about those resisting it. The pervasive deployment of datafication also enables inferences on those who opt-out, regardless of how techniques such as anonymisation are being applied[52].

Even when people are not chronically online, do not have social media apps, and largely stick to a more analog life, companies continue to collect data on them. In their book, *The Secret Life of Data: Navigating Hype and Uncertainty in the Age of Algorithmic Surveillance* Aram Sinnreich and Jesse Gilbert[53] call this data holes. Even a data hole is data. This is because of the interconnected nature between digital devices of people around them. This means that even gaps between data is meaningful data. Given this, datafication impacts the ways in which we interact with each other and our perception of what is private or public information.

---

[50] C. Véliz, *Privacy Is Power: Why and How You Should Take Back Control of Your Data*, Melville House, Brooklyn-London 2021.

[51] O.J. Gstrein, A. Beaulieu, *How to Protect Privacy in a Datafied Society? A Presentation of Multiple Legal and Conceptual Approaches*, in «Philosophy & Technology», XXXV, n. 1, 2022, p. 5.

[52] Ivi, p. 3.

[53] A. Sinnreich, J. Gilbert, *The Secret Life of Data: Navigating Hype and Uncertainty in the Age of Algorithmic Surveillance*, The MIT Press, Cambridge (MA) 2024.

While this may not alarm some consumers, it reflects a growing trend in mass surveillance in the United States, particularly in more subtler forms like digital surveillance. For this paper, I adopt a broad understanding of digital surveillance to refer to the use of technology by governments, third parties, corporations, or individuals to collect personal data, monitor, and track people, such as through biometric data, online communications, or networks. With this in mind, car manufacturers are collecting data across the board, which reports have found include consumers' sexual preferences, to their exact locations, such as the exact parking spot they parked in[54]. They do so through infotainment or driver-assistance systems, and Bluetooth. This digital surveillance raises concerns regarding consent, transparency, and the weaponization of data.

For example, a report by the Mozilla Foundation found that the amount of personal data collected by car manufacturers is already alarming. Even more concerning is that a staggering 56% of manufacturers disclosed that they share consumers' personal information with the government or local law enforcement, often without formal requests. As Mozilla's privacy team members explain in their consumer-facing research report, «A surprising number (56%) also say they can share your information with the government or law enforcement in response to a "request." Not a high bar court order, but something as easy as an "informal request"». When they are not selling consumer personal information to law enforcement or the government, they are selling it to data brokers. According to Brown University Office of Information Technology[55], data brokers are, «companies that collect, analyze, and sell large volumes of consumer information, often without direct consent, to third parties, for various purposes, such as targeted advertising and risk assessment» Data brokers will then sell the information they bought to other companies, law enforcement agencies, or anyone willing to buy that information.

Cars manufacturers are not the only ones to collect and then sell consumer data to data brokers. In fact, almost every company or entity we interact with will collect and sell out data at some point. This includes hospitals and social media websites like Facebook. Consequently, digital surveillance of women post-Roe v Wade increased and became a lucrative frontier for data brokers. Some of their highest bidders turned out to be anti-abortion groups. For instance, a *Wall Street Journal* report by Byron Tau[56] detailing that anti-abortion groups were using cellphone data to target people who visited Planned Parenthood. Following this report, Senator Ron Wyden initiated an investigation and found that the anti-abortion organization, The Veritas Society, hired advertising agency Recrue Media, who then used the data broker company Near Intelligence to obtain consumer location data from each reproductive health facility and their parking lots. The Veritas Society proudly proclaimed that just in Wisconsin they delivered 14.3 million ads

[54] A. Greenberg, *Subaru Security Flaws Exposed Its System for Tracking Millions of Cars*, in «Wired», 23 January 2025: https://www.wired.com/story/subaru-location-tracking-vulnerabilities/; O. Povey, *Concerns Raised over Tesla Spying on Drivers from Their Cars*, in «AS», 3 April 2025: https://en.as.com/latest_news/concerns-raised-over-telsa-spying-on-drivers-from-their-cars-n/; T. Klosowski, *How to Figure Out What Your Car Knows About You (and Opt Out of Sharing When You Can)*, Electronic Frontier Foundation, 15 March 2024: https://www.eff.org/deeplinks/2024/03/how-figure-out-what-your-car-knows-about-you-and-opt-out-sharing-when-you-can.

[55] Brown University Office of Information Technology, *Learn About Data Brokers*, 2025.

[56] B. Tau, *Antiabortion Group Used Cellphone Data to Target Ads to Planned Parenthood Visitors*, in «The Wall Street Journal», 18 May 2023: https://www.wsj.com/us-news/society/antiabortion-group-used-cellphone-data-to-target-ads-to-planned-parenthood-visitors-446c1212.

to people who visited abortion clinics[57]. Furthermore, Recrue Media's Cofounder admitted, «The company used Near to target ads to people who had visited 600 Planned Parenthood locations in the lower 48 states»[58].

To assess the ethics of such surveillance, I turn to Carissa Véliz's work on privacy and surveillance. Even when American women do not realize it, they are under significant amount of digital surveillance that may show up as something as simple as an anti-abortion ad that feels like it came out of the blue. In her chapter, "The Surveillance Delusion" in *The Oxford Handbook of Digital Ethics* she discusses surveillance delusion. Véliz believes that this normalization of the amount of surveillance we are under causes people to miscalculate the costs of surveillance. In particular, she believes that people fall into what she calls the surveillance delusion.

> Under surveillance delusion, only the benefits of surveillance are considered, and, as a result, surveillance is taken to be a convenient solution to problems that could be solved through less intrusive means—all without realizing that surveillance itself may be creating more weighty problems in the long run than the ones it is solving[59].

Consumers and the public struggle to understand the actual amount of harm that surveillance causes, especially when companies frame the technology in exciting or positive ways, like with autonomous vehicles. For example, people who buy Electric Vehicles (EVs) may do so wanting to be more environmentally conscientious or for their enhanced driver assistance systems, like in Tesla's. They most likely do not consider the amount of data the Tesla collects on them, including personal conversations they have inside the car. This is because having the ability to have a car that drives autonomously periodically offers a tradeoff with convenience that makes the surveillance component less important.

In her book, *The Ethics of Privacy and Surveillance*[60], Véliz discusses the duties that come with the use of surveillance. Véliz states:

> Whenever surveillance is implemented (even when it is justified), successive waves of duties arise from it, because duties to protect the right to privacy are being unfulfilled. Those new duties include informing the targets of surveillance (unless a criminal investigation necessitates temporary secrecy), keeping the data safe, deleting sensitive data as soon as possible, and minimizing the possible harms of surveillance[61].

This framing helps us understand the duties owed to those under surveillance. It is not always clear because of how common it is for companies not to obtain consent or mishandling personal or sensitive data. Take for example the case with the car

---

[57] R. Wyden, *Letter to FTC and SEC regarding Near Intelligence Inc.*, 13 February 2024: https://www.wyden.senate.gov/imo/media/doc/signed_near_letter_to_ftc_and_sec.pdf.

[58] R. Wyden, *Letter to FTC and SEC regarding Near Intelligence Inc.*, cit., p. 2; Z. McNeill, *Data Broker Sold Data From 600 Planned Parenthood Visits to Anti-Abortion Group*, in «Truthout», 15 February 2024: https://truthout.org/articles/data-broker-sold-data-from-600-planned-parenthood-visits-to-anti-abortion-group/.

[59] C. Véliz, *The Surveillance Delusion*, in C. Véliz (edited by), *The Oxford Handbook of Digital Ethics*, Oxford University Press, Oxford 2024, pp. 555-574: 556.

[60] C. Véliz, *The Ethics of Privacy and Surveillance*, Oxford University Press, Oxford 2024.

[61] Ivi, p. 139.

manufacturers. They collected data on consumers without informing them that their infotainment centers, Bluetooth, and voice-assistants collected and sold data to third parties. Law enforcement agencies and the government, were among the third parties that received consumer data, even when they did not make a formal request. Even though those car manufacturers technically did not do anything illegal, they failed their duties to protect their consumers in almost every way.

Véliz continues to critique the normalization of surveillance and argues that in order for its use to be morally permissible, it must be proportionate and necessary. Véliz states, "Given that surveillance has costs, it has to be *necessary* in that comparable beneficial results cannot be achieved by less intrusive or harmful methods. The moral concept of *proportionality* refers to a moral constraint on actions that cause harm."[62] In this case, if surveillance is necessary, the benefits of its use must outweigh any risks, wrongs, or harms it creates than the overall bad it creates. If it meets this condition, then it is proportionate and morally permissible. Similarly, Véliz notes, «Necessity is concerned with comparing what will happen if an act is done with what will happen if alternative acts are done that are also means of achieving the same end»[63]. If surveillance is necessary it is because all alternatives, including not using surveillance, fails to achieve the results or effectiveness. Veliz continues:

> The main idea is that any action that has moral costs (wrongs or harms) must have a convincing justification for it to be morally acceptable. For harms or wrongs to be justified, they must be done in an attempt to accomplish something good. If an action only creates a harm, and has no benefit, then it is not morally acceptable, as harms are something that ought to be avoided, other things being equal.[64]

This insight is relevant given the sheer amount of datafication in our lives and their associated harms. For instance, it is difficult to find data brokers who, following the Dobbs decision, increased surveillance of women because selling that data to anti-abortion organizations proved extremely profitable, as morally permissible. This contributed to groups like the Veritas Society to monitor and track random women who visited Planned Parenthoods. Furthermore, all of those women had no idea they were under surveillance because some random organization suspected them of getting abortions. However, the women could have been seeking other services other than abortion care, especially at Planned Parenthood. Planned Parenthoods provide free or low-cost Sexually Transmitted Infection (STI) testing, cancer screening, birth control, breast exams, and other preventative treatments and cancer screenings, mental health services, prenatal and postpartum services, vaccinations, and gender-affirming care, as well as testing and treating urinary tract infections (UTIs)[65]. The data broker company acting on behalf of the Veritas Society violated the privacy rights of the women they put under surveillance. This caused an immense harm, without any benefit.

As a result of this surveillance, the women received targeted anti-abortion ads. This despite the fact that they could have visited Planned Parenthood for other reasons. For

[62] Ivi, p. 138.
[63] Ivi, p. 139.
[64] *Ibidem*.
[65] Planned Parenthood, *Our Services*, 2025: https://www.plannedparenthood.org/get-care/our-services.

instance, they may have gone because they suspected they had endometriosis, or found a lump in a breast, or needed an STI screening because they had a new partner. Instead, a company willing to abuse privacy laws holds their information and stores it indefinitely. If the Veritas Society's true goal was to send targeted ads to women of a certain age and demographic, they could have used other well-known and used alternatives. For example, other companies regularly use Meta or TikTok Ads Manager because they are less invasive and do not violate abuse privacy laws, while achieving the same results. These methods show the surveillance capability a single data broker took on behalf of one anti-abortion organization to send targeted ads on social media. Their use of surveillance was clearly neither proportional nor necessary, and thus morally impermissible. In fact, it sets a dangerous precedent for other companies and organizations to do the same, and against other vulnerable populations.

Now, imagine if Veritas Society was one of the companies that car manufacturers sold sensitive location data, too, as well, in addition to what they collected through Near Intelligence. Remember, car manufacturers already plan on installing these emerging technologies and their data collection features into newer models. These cars are likely the ones that will be operational during the transitional period Sparrow and Howard predict occurs before the ban. Nevertheless, it is important to note that their proposal is still speculative. This means, that even before their proposed ban, people will not be able to opt-out of the ban, so to speak. Their old cars will be able to sit in their garages or on the side of the street, but car manufacturers will and can disable the car, prohibiting any access to the roads. Similarly, the car manufacturer can and most likely will provide law enforcement with data on any human driver who defies the new ban. The caveat is that implementing and enforcing such a ban would require even greater levels of digital surveillance. Thus, I ultimately reject Sparrow and Howard's proposal on the grounds that it would be morally impermissible for the government to outlaw human drivers. This is because the level of surveillance required would be neither proportional nor necessary. I further support this stance by first presenting a troubling empirical case, followed by a hypothetical case study involving Saoirse.

In addition, the outlawing of human drivers is likely to receive significant pushback. If the sustained resistance to Seat Belt Laws since the 1980s is any indication of the public's response to laws they perceive as overstepping, then it is hard to imagine that the public would not react similarly, if not more zealously, to increased surveillance, surveillance delusion or not. Furthermore, it is difficult to imagine that a full outright ban on human drivers would go well, either. If a significant portion of the public continues to rally against seat belts, it is likely they will rally against increased surveillance and an outright ban on their right to drive. Similarly, increased surveillance and data collection are likely to face legal and democratic resistance and barriers, especially as public awareness about the data broker industry increases, as is evidenced by Senator Ron Wyden's letter to the FTC and John Oliver's humorous, but poignant, segment on data brokers[66].

[66] Last Week Tonight, *Data Brokers: Last Week Tonight with John Oliver (HBO)*, 11 April 2022: https://www.youtube.com/watch?v=wqn3gR1WTcA.

6. The Convergence of Restrictive Abortion Laws and Autonomous Vehicle Regulation in the U.S.

a. The Benefits and Perils of Banning Human Drivers

One such emerging technology is likely autonomous vehicles because they are the perfect amalgamation of mobility regulation, legal frameworks, and digital surveillance that overlap in complex ways that may draw people out of their surveillance delusion. In this section, I develop a critical analysis of the convergence of the previously-discussed abortion travel bans and proposed autonomous vehicle regulation in the U.S. From Sparrow and Howard's (2017) techno-optimistic perspective[67], the future of autonomous vehicle regulation, contains a list of potential benefits of banning human drivers and increasing the availability of driverless vehicles that outweighs any harms that may arise.[68] Their list of benefits includes a total shift in insurance and liability costs, from the human driver to the programmer or engineer; an increase in accessibility for previously excluded populations like the elderly or those with medical conditions; and an increase in efficiency and improvement of current public transportation systems. I do not disagree or dispute their listed benefits. Additionally, because users will need to input point A and point B of their travel destinations, there are some further benefits beyond Sparrow and Howard's list. They include prohibiting:

- Domestic abusers from violating temporary protection orders (TPO) or going anywhere near the victim;
- Convicted pedophiles or sex offenders from accessing known children-oriented places, like schools, theme parks, and homes with children;
- Reckless driving;
- Driving under the influence (DUI).

The list of potential benefits contains various reasons to support a ban on human driving. The ban on human driving could theoretically reduce traffic fatalities, kidnappings, and instances of domestic violence. It could also lead to local and federal governments prioritizing public transit, making public transit more accessible and effective. Additionally, this technology could increase independence for the elderly and those previously under driver restrictions due to medical reasons. While these benefits are promising, it is important to keep in mind some potential perils.

Potential Perils:

---

[67] In his article, "Techno-optimism: an Analysis, an Evaluation and a Modest Defence", John Danaher (J. Danaher, *Techno-Optimism: An Analysis, an Evaluation and a Modest Defence*, in «Philosophy & Technology», XXXV, n. 2, 2022, p. 11: https://doi.org/10.1007/s13347-022-00550-2) states, "This article focuses on impersonal forms of optimism. Techno-optimism is the stance that holds that technology, defined here in largely material and instrumentalist terms, plays a key role in ensuring that the good does or will prevail over the bad." Danaher's definition points to a general understanding of emerging technologies largely producing benefits that outweigh any harms generated by its adoption. Danaher offers a moderate defense of techno-optimism. Danaher (ivi, p. 27) states, «Despite this, however, I have concluded that a modest form of techno-optimism, one that does not assume that technology will save humanity by itself, nor that technology is sufficient for the good to prevail, is defensible» Danaher's modest defense does point to more of a viewpoint that this paper takes: if we establish the right regulatory frameworks and policies, then technology has the potential to improve humanity and allow good to prevail over bad.

[68] To view Sparrow and Howard's list see R.J. Sparrow, M. Howard, *When Human Beings Are like Drunk Robots*, cit., pp. 212-213.

- Misuse and abuse of current policy frameworks by policy and lawmakers.
- Surveillance of women and marginalized communities through using:
  - Digital surveillance;
  - Data brokers;
  - Social media;
  - Geolocation.
- Limitations of individual freedom of movement.
- Automated Incident Reporting and the Militarization of US Police
- Interpretations of regulations in relation to other laws and policies.

Although we have strong reasons to support the ban of human drivers, the perils do point to nontrivial ethical problems when applied in tandem with other regulations. Like previously discussed, vehicles have unprecedented abilities to surveil people inside and around them, and then selling that information to data brokers, marketing firms, and the government (Ferris 2024) without much prompting or justification. As digital surveillance advances in the United States, it is plausible that authorities will likely increase monitoring not only women, but also other heavily surveilled populations such as political dissidents, immigrants, and historically marginalized and minority communities. Consequently, digital privacy could become a more pressing issue as current vehicles continue to become more advanced and regulations continue to remain lax.

b. Social Media as a Digital Canary Bird

Social media users are similarly vulnerable to the misuse, selling of, and abuse of their personal data, due to ambiguous regulations around digital privacy. They suffer the consequences of techno-optimism and lack of protections in the name of innovation. If Sparrow and Howard's own policy proposal to ban human drivers comes to fruition, there is a potential for there to be the perfect dystopian combination of complementary, but destructive policies.

This is where the laws around driverless vehicle regulations and restrictive abortion laws may converge. In the near future, we could live in a time period where a ban of human drivers exists, which would be in a post-Roe v Wade world. Due to the weak regulatory landscape surrounding digital privacy, the overturning of Roe v Wade and the Chevron doctrine, in the U.S. could subject people of reproductive age to heightened surveillance from their social media to the driverless vehicle they use to travel.

c. Current Empirical Cases

For women of reproductive age in the US, they are under more surveillance than they realize, especially in areas with strict abortion laws that allow and have reporting systems in place (e.g., Texas). For instance, due to resistance from the courts and the public, anti-abortion groups have enlisted the help of former "sex partners" of women who get an abortion or are suspected of getting an out-of-state abortion[69]. For example, the anti-

[69] C. Kitchener, *Antiabortion Advocates Look for Men to Report Their Partners' Abortions*, in «The Washington Post», 17 January 2025: https://www.washingtonpost.com/investigations/2025/01/17/texas-abortion-pills-lawsuit/.

abortion group, Texas Right to Life, uses social media campaigns to recruit aggrieved current or former male partners of women who they suspect of obtaining an abortion, planning to, or who have obtained an abortion. The group became inspired by anti-abortion lawyer, Jonathan Mitchell's use of legal loopholes to file lawsuits against people who assist women seeking abortion care, through using Texas's wrongful death statute. In her article, "Antiabortion advocates look for men to report their partners' abortions" Caroline Kitchener reports:

> For male partners initiating a wrongful-death case, the first step is frequently to request highly personal and sensitive information from the woman who chose to end her pregnancy, compelling her to hand over text messages and other documentation related to her abortion that could then be made public during discovery.[70]

Anti-abortion groups have not stopped at recruiting men to report their partners after either coercing or deceiving women into providing personal information to them. Texas's restrictive abortion laws have also emboldened police officers in Texas to stalk women they suspect may have had an out-of-state abortion in a state where abortion is legal. On June 9th, 2022, The Illinois secretary of state had to initiate an investigation into a suburban Chicago police department sharing data from an automatic license-plate reader to a Texas sheriff deputy. This discovery occurred due to a website, 404 media, reporting that the Texas sheriff sent out a nationwide request for data that amounted to requesting access from roughly 83,000 cameras operated by Flock Safety[71].

The Texas sheriff justified violating state laws by stating he sought the digital surveillance data because the woman's family reported that they suspected the woman had obtained an abortion in Chicago. The woman did not violate or break the law in Illinois, where abortion is legal, but her data from license plates readers in the state were still collected and given to the Texas sheriff anyway. These two cases are just two of many likely unreported efforts of anti-abortion groups to weaponize legal loopholes, judicial precedents, and digital surveillance technologies to persecute women seeking abortion care, even in states where abortion is legal. This type of weaponization, coupled with the growing interoperability of digital surveillance systems, sets the stage for increasingly aggressive and systematic attempts to restrict mobility. Even without driverless car bans, these cases illustrate real-world scenarios in which legal loopholes, interjurisdictional data-sharing, and surveillance technologies are already in use.

d. A Collision of Regulations Case-Saoirse

To help concretize what a convergence of restrictive abortion laws and a ban on human drivers could look like in practice, here is a case study.

---

[70] *Ibidem.*

[71] J. O'Connor, *Illinois Investigates Police for Sharing License Plate Data with Texas Sheriff*, in «AP News», 12 June 2025: https://apnews.com/article/abortion-access-immigration-license-plate-readers-surveillance-13fac7c045df3c5e5145f6d4e4c4db28; A. Mahdawi, *A Dystopian Surveillance Fear Has Become Reality in Texas*, in «The Guardian», 31 May 2025: https://www.theguardian.com/commentisfree/2025/may/31/a-dystopian-surveillance-fear-has-become-reality-in-texas.

Some Key points:

- State A-abortion is illegal;
- State B- abortion is legal;
- Driverless vehicles exist;
- Human Drivers are Banned.

Case:

According to Sparrow and Howard's ideal regulatory framework and anti-abortion groups ideal abortion regulations, resides Saoirse. Saoirse is a 24 year-old woman, who is 12 weeks pregnant. She lives in State A. She wants to get an abortion from State B. However, humans cannot drive. The only type of available transportation that is an individual-public driverless vehicle or public driverless fleets. Saoirse cannot drive herself to an abortion clinic in State B because she lives in State A, which prohibits such inputs. In addition, all users must first input their information in order to access a driverless vehicle.

Saoirse uses a period tracker app and the app sold her health data to data brokers. Her profile now indicates that she is of reproductive age and possibly pregnant due to missed periods. Due to the possibility of being pregnant, through OTA systems, the app will not allow her to schedule a pick-up or drop-off near State B. Instead, it indicates permitted roads and routes pregnant people can use, as a pregnant person living in and subject to laws and regulations in State A. Saoirse knows not to request help via social media or messaging apps. In addition, there is a possibility that the driverless vehicle she uses collects and sends her information to local law enforcement. Thus, due to the heavy surveillance the two overlapping laws allow, Saoirse cannot access abortion care or support. Furthermore, the convergence of the regulations, restrict her mobility, especially interstate mobility, until further notice.

e. Upshots

In understanding Saoirse's case study, it illustrates the perils of when Sparrow and Howard's regulatory framework converges with in-practice anti-abortion groups ideal regulations. Sparrow and Howard's techno optimism fails to account for when their ban on human drivers would converge with existing technological and constitutional and legal precedents, and policy frameworks. Additionally, the case study expands to include other heavily-surveilled populations, like political dissidents, immigrants, and other marginalized and minority communities. Moreover, the case study shows the potential for driverless vehicles to play a fundamental role in the surveillance state and either reducing or undermining bodily autonomy and freedom of mobility for millions of Americans. Thus, although we have strong reasons to support banning human drivers, such policies need to take into consideration the interplay with existing regulations and their ethical implications. When we consider the ideal outcomes of these regulatory proposals according to their sponsors and advocates, that is when we can see the true potential perils. Moreover, Saoirse's case study helps explore broader policy implications regarding policy and the often overlooked impact on marginalized communities.

As the examples illustrate, state surveillance currently uses various tools and resources (e.g., digital borders and surveillance) and those are likely to increase in scope and scale in the near future. These also include taking advantage of interoperability between government agencies and data systems, a weak digital regulatory landscape, and the use of data brokers to scrape and aggregate user information. The banning of drivers increases the use of digital profiles for travelers accessing driverless fleets. In addition, proposed surveillance measures to enforce restrictive abortion laws, could further exacerbate state surveillance. Together, these converging regulations could potentially reduce or undermine the bodily autonomy and freedom of mobility of millions of Americans. However, the state would need access to an unprecedented amount of surveillance technologies beyond what they already utilize. Although I am uncertain how seriously the United States will approach digital regulations, following the EU's suit is one potential way to mitigate the risks I have addressed throughout the paper.

7. Conclusion

Throughout the paper, I centered my discussion on the convergence of restrictive abortion laws and the banning of human drivers. I examined the possibility of banning human drivers and what a driverless future looks according to Sparrow and Howard, and the transition to such a future. I further provided an examination of current and proposed restrictive abortion laws in the United States and their future policy implications, outlining the various approaches anti-abortion organizations are taking to circumvent constitutional challenges to their goal to ban abortion travel across the United States. Afterwards, I examined current forms of digital surveillance and mobility monitoring the United States already engages in, the current administration's expansion of those capabilities, and the potential implications. Then, I presented a case study to show the ethical implications of the convergence of autonomous vehicle regulation and restrictive abortion laws. This paper shows how discussions about self-driving car regulations and current restrictive abortion laws in the United States, overlap. It highlights the serious implications for bodily autonomy, freedom of mobility, and surveillance, particularly for women and marginalized communities.

However, despite an increase in digital surveillance and other restrictions, the public has continued to show significant resistance and pushback that has meaningfully stalled or made it more difficult for these more restrictive or despotic policies to pass. It is crucial that the public sustains this resistance. Equally vital is the ongoing efforts by academics, journalists, and other organizations to continue sharing knowledge and information. My hope is that by showing the potential harms of these overlapping policies, I can contribute to a broader public discourse. The goal is not to spark moral panic, but rather to foster greater awareness of the potential consequences when these policies meet.

Beyond automation: the essential role of librarians in the age of generative AI[a]

Pier Francesco Micciché[*]

Abstract

Gli LLM sono ampiamente considerati strumenti dal potenziale enorme, spesso destinati a sostituire i lavoratori umani in vari settori, compreso, nel prossimo futuro, quello dei bibliotecari. Al contrario, gli LLM hanno più che mai bisogno di professionisti dell'informazione umani che aiutino la società a valutarne i risultati, comprenderne il funzionamento e riconoscerne gli errori e i limiti. I bibliotecari possiedono le competenze necessarie per valutare criticamente i risultati dell'IA, mitigare i pregiudizi e promuoverne un uso responsabile. Questo documento esplora l'evoluzione del rapporto tra biblioteche e LLM, sfidando l'idea errata che essi siano i principali concorrenti nella gestione delle informazioni. Al contrario, le biblioteche possono sfruttare l'IA per migliorare i propri servizi, posizionandosi come centri nevralgici per l'alfabetizzazione all'IA. Educando gli utenti sui limiti dell'IA, sulle questioni etiche e sui potenziali rischi di disinformazione, i bibliotecari possono promuovere un approccio critico a queste tecnologie. In definitiva, questo documento sostiene che l'integrazione dell'IA negli ecosistemi della conoscenza non deve emarginare i bibliotecari, ma piuttosto rafforzarne il ruolo di educatori e mediatori critici.

Parole Chiave: alfabetizzazione all'IA, biblioteche, bias dell'IA, allucinazioni.

LLMs are widely viewed as tools with vast potential, often predicted to replace human workers in various fields including, in the near future, that of the librarian. On the contrary, LLMs need human information professionals more than ever to help society evaluate its results, understand how it works, and recognize its mistakes and limitations. Librarians possess the expertise to critically assess AI outputs, mitigate biases, and promote its responsible use. This paper explores the evolving relationship between libraries and LLMs, challenging the misconception that they are their main competitors in information management. Instead, libraries can leverage AI to enhance their services while positioning themselves as key hubs for AI literacy. By educating users on AI's limitations, ethical concerns, and potential misinformation risks, librarians can foster critical engagement with these technologies. Ultimately, this paper argues that AI's integration into knowledge ecosystems must not sideline librarians, but rather empower them as educators and irreplaceable critical mediators.

[a] Received on 12/04/2025 and published on 09/12/2025.
[*] Università del Piemonte Orientale – Convenzione FINO, e-mail: pierfrancesco.micciche@uniupo.it.

Keywords: AI literacy, libraries, AI bias, hallucinations.

1. Introduction

In recent years, Generative Artificial Intelligence (henceforth: GAI) tools have increasingly spread across domestic, commercial, and public domains. This rapid expansion opens up new opportunities but also introduces significant ethical, cultural, and social risks. On the one hand, the increasing sophistication of these tools appears to threaten numerous professional roles, including that of the librarian. On the other hand, risks related to an incorrect or unconscious use of AI may include the dissemination of fake news, ethical misuse in scholastic or academic contexts, and the reproduction of errors and biases absorbed by the dataset of LLMs. In this sense, rather than representing the weak party, librarians are uniquely positioned to become essential allies and play an important role in shaping a conscious use of such powerful tools. Their skills in fact-checking, bias checking, scrupulous checking of the reliability of sources and information retrieval become even more important when it comes to critically weighing the activity and output of GAI softwares. Furthermore, although such a development of AI could not have been foreseen when the Sustainable Development Goals[1] (SDGs) were enacted, among the goals of the United Nations 2030 Agenda, we could affirm that this vision goes in the same direction as goals n.4 - quality education, and n.9 - industry, innovation and infrastructure. Libraries are one of the main headmasters for lifelong education and life-long learning, especially in the western countries, and their role in Goal 4 of the UN 2030 Agenda seems beyond question. At the same time, librarians (especially academic ones) can benefit from the use of AI-based platforms for retrieving scholarly resources.

In the following sections, we will explore the relationship between libraries and AI and vice versa, by weakening the idea that they are enemies of one another competing in the realm of information retrieval. We will then discuss how GAI can benefit from the existence of libraries, and how libraries can benefit from the use of AI. Finally, we will show the gap that separates AI from a good librarian, and why their role remains irreplaceable to this day.

2. Libraries and AI. Enemies or allies?

What is generative AI? The answer is not as simple as one might believe. Even if a full framework is beyond the aim of this article, we need to know at least *what* GAI *is not*. The application of the attribute of 'intelligent' to AI, for instance, is disputed[2]. Indeed, it is necessary to ask whether intelligence has to do only with *planning*, *learning*, *adapting* and *interacting* or also with *awareness*, *consciousness*, *emotional* and *moral life*. Most scholars, albeit for

[1] United Nations, *Resolution adopted by the General Assembly on 25 September 2015. Transforming our world: the 2030 Agenda for Sustainable Development A/RES/70/1*, 2015, https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf.

[2] See, for instance: F. Chollet, *On the Measure of Intelligence* in «arXiv», 2019, doi:10.48550/arXiv.1911.01547 and F.M. Cianciaruso, *L'intelligenza artificiale è realmente intelligente? Un percorso tra embodiment e macchine linguistiche* in «DILEF», IV, n.4, 2024, doi: 10.35948/DILEF/2024.4359.

different reasons, do not believe in the full application of the concept of 'intelligence' to AI tools.

Pezzulo et al.[3], Manyika[4] and Alombert[5], for instance, think that GAI simulates tasks but lacks genuine understanding. Cicero[6], together with García Peñalvo[7] et al. share the idea that GAI excels in rule-based and mimetic tasks yet falls short in creativity, autonomy, and subjective experience. Yet, according to Cianciaruso[8], the watertight bulkhead that separates formal computational processes from a dimension of meaning is abolished by embodiment, that is, by the embedding of mental content – and thus of its decoding – in the human sensorimotor system. This is because «the data sets on which the machine performs the training processes, are the product of the culture of human agents, and, therefore, embodied»[9]. GAI would thus be a more than syntactic intelligence that is not limited to symbol manipulation. In other words, if it is true that *nihil est in intellectu quod prius non fuerit in sensu*, and datasets were trained with data from creatures endowed with sentience, then that of AI is an *intellectus* in its proper sense. According to embodied cognitive sciences:

> The selection of the data and examples from which the machine is trained are in fact external to the machine itself, as they are the result of an external choice, prior to the training, and, therefore, to the implementation of the neural model[10].

However, we may doubt that indirect, third-person embodiment can be sufficient, as well as that all datasets are always the result of sensations of creatures with adequate sensorimotor apparatus. If we adhere to this last conception, we must admit that indifference to the semantic contents of inputs and outputs, and especially the indifference to truth (understood in the Aristotelian sense of '*adaequatio intellectus et rei*'), still mark a distinction between human and machine intelligence, and between the two concepts of "understanding". Moreover, although embodiment implies continuity between human and artificial neural architecture, the alleged rootedness would remain parasitic to the human agent and lack autonomy. Conversely, if we adhere to the opposite position, we will instead have a harder time recognizing this intelligence as artificial, unless we admit that it is an artificial intelligence applied on human mental content (i.e. rooted in human motor-sense system).

---

[3] G. Pezzulo, T. Parr, P. Cisek, A. Clark, K.J. Friston, *Generating meaning: active inference and the scope and limits of passive AI*, «Trends in Cognitive Sciences», 2023, doi: 10.1016/j.tics.2023.10.002, pp. 97-112.

[4] J. Manyika, *Getting AI Right: Introductory Notes on AI & Society*, «Daedalus», CLI, n. 2, 2022, doi: 10.1162/daed_e_01897, pp. 5-27.

[5] A. Alombert, *Panser la bêtise artificielle. Organologie et pharmacologie des automates computationnels* in «Appareil», XXVI, 2023, doi: 10.4000/appareil.6979, pp. 1-18.

[6] F. Cicero, *L'italiano delle intelligenze artificiali generative*, in «Italiano LinguaDue», XV, n,2, doi: 10.54103/2037-3597/21990, pp. 733-761.

[7] F.G. García Peñalvo, F. Llorens-Largo, J. Vidal, *La nueva realidad de la educación ante los avances de la inteligencia artificial generativa* in «RIED: Revista Iberoamericana de Educación a Distancia», XXVII, n. 1, 2024, doi: 10.5944/ried.27.1.37716, pp. 9-39.

[8] F.M. Cianciaruso, *L'intelligenza artificiale è realmente intelligente?* cit., p. 4.

[9] Ivi, p. 9.

[10] *Ibidem*.

Similar considerations can be applied to the appellation of 'oracle' to GAI[11]. Even if our perception of use may resemble that of someone questioning an oracle, this metaphor can lead us to mistakes. If an oracle has a *superhuman ability* to know *human things* (i.e. a knowledge rooted in the sensorimotor system), analogy works (even if, in order to do this, he would have to know an endless number of things, and likely have had a very long life). If, however, an oracle is someone divinely inspired, that is, he reports, as if in ecstatic possession, things he has not experienced himself, GAI tools could not be properly called 'oracles'. GAI possesses and reworks, even creatively, knowledge that it has not personally experienced, but that some human being has likely experienced, and therefore GAI it does not possess this information as a divine gift. In other words, its frightening "intelligence" (and, as Roncaglia observes, creativity[12]) derives from the *reworking and connection of elements* that probably we had not thought to process or link together. This is far from insignificant, but is super-human more than beyond-human[13]. Moreover, an oracle knows the truth and gives it to the listener through riddles; GAI gives clear answers that can be made even simpler, but they don't really know the truth.

Neither the definition of 'encyclopedia' as a structured, homogeneous, and verified repository of information, seems good for GAI, not so much because of the amount of data, but because of how they are linked and how they are provided to the user. Rather, it is the datasets that the GAI draws on that are (potentially) similar to encyclopaedias.

Finally, even the use of the term 'calculator' could be problematic in some cases, especially when understood in a narrow sense. The principles on which LLMs are programmed are best defined as an integration of mathematical functions and probabilistic (usually not logical) algorithms that enable the simulation of cognitive capabilities such as machine learning, rather than mathematical problem solving. As Manna notes, in science majors such as chemistry, mathematics and physics,

> ChatGPT often gets the answers wrong: although it may seem counterintuitive, services like ChatGPT in fact do not do the math right, because they do not reason logically and mathematically, but generate their answers according to probabilistic criteria. In these cases, for algebraic exercises some students rely on other software (some actually available long ago), such as Photomath: an app that automatically does math exercises from a picture of a function written in the notebook or on the blackboard[14].

Of course, as Roncaglia wrote:

> the results obtained by ChatGPT in tasks that require complex updates and evaluations, such as working on mathematical or logical proofs, improve greatly if the system undergoes specific training in breaking down complex tasks into a sequence of simpler tasks[15].

---

[11] G. Roncaglia, *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*, Laterza, Bari-Roma, 2023.

[12] See *Ibidem*, p. 107.

[13] I mean that GAI is faster than any human being and smarter than any individual human being, but what it is capable of doing remains human in its methods. It is not "inexplicable".

[14] N. Manna, *ChatGPT sta cambiando la scuola* in «Il Post», 31/03/2025, https://www.ilpost.it/2025/03/31/chatgpt-intelligenza-artificiale-scuola/?utm_medium=social&utm_source=facebook&utm_campaign=lancio.

[15] G. Roncaglia, *L'architetto e l'oracolo*, cit., p. 99.

as it has been doing automatically for some time now.

So, GAI is not able to be called easily, and without problematizing, neither an intelligence (and in a certain sense not even an artificial one! - consider the embodiment of their data sets), nor an oracle, nor an encyclopedia, nor a calculator.

Yet, what we are here more interested in investigating is what we can reasonably expect AI to help us with and what it does not, and how our approach to it can make it as useful and less dangerous as possible. According to Walter Riviera, italian AI Engineer and expert, we will define these tools as «excellent text processors [...] who do not give the most correct or sensible answer but the most probable one»[16]. For this reason, we can trust the tool for

> All tasks that have to do with syntax, i.e. text manipulation, summaries, translations, alternative sentences to express the same thing, writing an official e-mail to the boss...for these things we can feel quite comfortable. When the need is on the content, on the semantics, on the meaning of what we are expressing...maybe I would give it a read before sending the e-mail[17].

It is no chance that this is the use that large software platforms such as Adobe, Google, and Amazon have made of it: the first one with pdf documents, the second one with e-mails, the third one with customer reviews of products. Large quantities of sentences, even in different languages or codes, synthetised for a broad understanding that optimises the reader's time. This is exactly what S.R. Ranganathan, one of the founders of the modern library science, called for in the fourth of his famous five laws of library science ("Save the time of the reader"[18]). Thus, libraries too, especially digital ones, dealing with large amounts of information and text, can benefit from GAI. If we consider how the role of manipulating texts is often performed admirably by GAI, many information-related professions seem to be seriously endangered by the emergence of these platforms. Among these is certainly that of the librarian, understood above all as a mediator between the user and the (physical or digital) document, and as a cataloguer clerk.

As was already written elsewhere[19], different AI software can support librarians roughly in the performance of nine tasks, also through the use of specific plugins:

1. Automation of cataloguing and indexing, creation of metadata, abstracts and keywords, error correction; more generally, acceleration of mechanical entries and processes;
2. Communication and marketing (producing from scratch, or revising newsletters, graphics, claims, copy, logos, social media posts, press releases...);
3. Personalised recommendations and reading suggestions for users from the library catalogue (using machine learning algorithms to analyse reading behaviour, keyword

[16] W. Riviera, *Quando sbaglia l'intelligenza artificiale? Abbiamo messo alla prova ChatGPT e Deepseek: gli errori*, https://www.youtube.com/watch?v=89C0jCgtBuk, 27/02/2025.
[17] *Ibidem*.
[18] S.R. Ranganathan, *The Five Laws of Library Science*, Madras Library Association, Chennai 1931.
[19] M. Capozza, M. Mander, P.F. Micciché, F. Viazzi, *Intelligenza umana e artificiale: le biblioteche pubbliche creatrici di una tela in divenire* in Associazione Biblioteche Oggi (edited by), *Biblioteche Oltre. I nuovi territori dell'interdisciplinarità (Atti del Convegno delle Stelline 2025)*, Bibliografica, Milano 2025, doi: 10.53134/9788893576765-311, pp. 311-323: 317.

extraction and association between related readers), including ChatBots and virtual voice assistants;
4. Digitisation of paper documents and optical recognition of alphanumeric characters (Optical Character Recognition, OCR), shapes, signs, notes or neumes hard or slow to decode;
5. Information retrieval: finding papers, monographs, documents on a certain topic;
6. Speech synthesis and tools for accessibility;
7. Analysis of data, reading behaviour, trends, demand forecast from Library Management Systems (LMSs);
8. Advice for the librarian about the organisation of events in the library, the management and maintenance of equipment, and the arrangement of space and furniture (better arrangement of collections in shelves and shelves in physical space);
9. Gamification, augmented virtual reality experiences, customised learning paths.[20]

The list seems to encompass almost all the tasks of the librarian, legitimising a certain fear about the future of the profession. Nevertheless, so far the application of AI to library work has been rather limited. The fact that platforms can potentially help with all these tasks, in fact, does not mean that they can do so without the need of any human supervision, nor that they can always do it better or faster than any human worker. Relieving librarians of the more mechanical and repetitive tasks in order to leave them with the more social and 'human' tasks is a goal that has been pursued for several years now with multiple Information Technologies (IT) tools (RFID - Radio Frequency Identification tag and readers, electronic trolleys, Library Management Systems, self-checkout stations...). The same purpose drives the use of GAI in libraries, allowing staff to focus on tasks requiring human creativity and supports strategic decision-making with actionable insights[21]. Nevertheless, this has made the human irreplaceability of the librarian even more evident where he or she has been able to express his or her full potential. The reason for that is not only that machine operations are always liable to mistakes, but that libraries are not simply repositories of books to be kept.

Even apparently mechanical tasks, moreover, are only so to a certain extent. As Ciocci et al. note, «keyword production is thus, by its very nature, an activity of interpretation»[22]. It is not enough to count how many times a certain word appears in the text to state with certainty the subject matter of a document. Even the creation of metadata, ontologies and classification systems, although often facilitated by software, is therefore not a totally-neutral operation, nor one that can be relegated to machines without human supervision. Conversely, librarians, although much slower in reaction time, are equipped with the specific discerning skills needed to structure the architecture of the datasets from which the processor draws. If libraries can make use of AI, in short, it would be much more useful to make use of libraries for those who want to consciously use AI.

[20] See F. Boateng, *The transformative potential of Generative AI in academic library access services: Opportunities and challenges*, «Information Services and Use», XLV, n. 1-2, 2025, doi: 10.1177/18758789251332800, pp. 140-147.

[21] Ivi, p. 140.

[22] D. Ciocci, M. Squarcione, F. Viazzi, *Biblioteche e AI Literacy: alcune riflessioni sulle nuove competenze per la mediazione informativa* in «Biblioteche Oggi Trends», X (2), 2024, doi: 10.3302/2421-3810-202402-076-01, pp. 73-84.

But the relationship between libraries and AI is not limited to performing mechanical tasks faster and more efficiently. Public and academic libraries can assume the role of antibodies - better, vaccines - to an uncritical use of AI. This may happen by AI literacy courses, showing the limitations of conducting purely AI-based source research and the benefits of using human eyes and librarian expertise, especially in sensitive and specific situations. But also through warning against the underrepresentation of categories and exceptions that are considered negligible but are not so for science. Furthermore, librarians trained in AI can easily show how some of the answers considered 'original' by GAI are not really so (even if it is "in good faith"), and it is therefore dangerous to delegate certain tasks to them without checking. In an interview with eight library employees, all respondents «were aware of so-called hallucination problem in generative AI (generation of nonsensical or untrue content) and biases embedded in these systems, but stressed that libraries should learn about this technology and use it to their benefit»[23].

GAI limits are manifested both with respect to the limits of the machine as a machine, and also as a platform operated by large corporations interested in profit. According to Parisi & Dixon-Roman, the ethical violations of GAI are inherent to the economic ideology underlying the development of these technologies[24].

Libraries, in short, can and hopefully must become hubs for AI literacy for ordinary people, and in particular, for those who want to use these tools to research, produce or even just get new knowledge. Librarians can share awareness of the mechanisms underlying AI and their implications for everyday and scientific use, and this is their added value.

3. All that glitters is not gold. Intrinsic and extrinsic limits of Generative AI

Among the operations that GAI cannot do, some cannot be performed simply because it is a machine and not a human, and certain because it is a machine programmed in a certain way rather than another. These limitations reveal what we can reasonably expect from LLMs and especially because, as we already said, it is not an oracle.

Alongside the list of what AI can do for libraries, we will therefore have to draw up a list of what it cannot do, and is unlikely ever to be able to do. Although the following list is not exhaustive, it would include:

1. Setting goals for oneself;
2. Acting and "reasoning" without prompting;
3. Access to anything that cannot be converted into digital data (such as sensations themselves: holding a book, leafing through it, enjoying it, etc.);
4. Have an original and "felt" ethical judgment;
5. Empathy;
6. Work beyond minimum data units.

---

[23] M. Hosseini, K. Holmes, *The Evolution of Library Workplaces and Workflows via Generative AI*, «College and Research libraries», LXXXIV, n.6, 2023, Url: https://crl.acrl.org/index.php/crl/article/view/26094/34016?utm_source=chatgpt.com https://blog.library.gsu.edu/2024/12/20/librarians-shape-ethical-research-in-an-ai-driven-world-23-24-impact-report/?utm_source=chatgpt.com.

[24] L. Parisi & E. J. Dixon-Roman, *Data capitalism, sociogenic prediction, and recursive indeterminacies*, in P. Mörtenböck & H. Mooshammer (dds.), *Data Publics Public Plurality in an Era of Data Determinacy*, Abingdon (UK), Routledge. 2020, doi: 10.4324/9780429196515-4.

About this last point, we must consider that, as a text string processor, most LLMs works through tokenization, and so are unable to recognize what is at a higher level of specificity than the token itself. In cases where tokens correspond to syllables or clusters of letters, even sophisticated tools such as ChatGPT may become incapable of formulating anagrams, alphabetizing similar words, recognizing the position of letters within a word[25].

Moreover, AIs' operation is more like that of a probabilistic forecaster than a supercomputer. If we ask most AI tools to generate an image of a clock showing 5:35 a.m., most of them will be unable to produce a right corresponding representation, while being able to instruct (verbally) someone on how they should draw it. This happens simply because most of the analog clocks "learned" from the dataset indicate other times (especially 10:10 a.m./p.m. like in Figure 1).

If we provide an LLM with the complete works of a certain author, it could certainly give us a list of the most frequently occurring words, but that does not mean it could determine whether or not they would be useful for our research. The word 'music', for example, never appears in Dante's entire *Divine Comedy*, yet there are numerous references to music, described using metaphorical expressions, and a lot of books were written about the fascinating use of music images in Dante's *Commedia*. A librarian knows this. For the same reason, GAI can pass off as original what is not, giving rise to copyright issues. If we ask a GAI to generate an

Figure 1: AI-generated pictures (Dall-E 2).

original story, it is very likely that it will give us existing stories, or pieces of them. The risk of unintentional plagiarism is so very high.

Thus, Daniela Tafani correctly observes that one cannot speak of 'hallucinations' for what are 'outputs of software architecturally incapable of distinguishing truth from falsity'. This calls for a shift in terminology: what is often called a 'hallucination' may simply

[25] At the time of writing, the free version of DeepSeek and Chat GPT are incapable of doing so, while Gemini succeeds.

be the predictable outcome of a probabilistic system devoid of truth-evaluating mechanisms[26].
If the control of these tools were not in the hands of large and rich companies, we could hope that the liberation of people from the most mechanical and repetitive jobs, and their reliance on intelligent machines, would become a means of emancipation for mankind, ensuring that humans do not behave like machines and machines like humans. At least not for the time being, and therefore Waelen considers AI ethics as a form of critical theory[27].

The highest risk, especially in the public administration, is that of overestimating the potential and possible uses of AI by only increasing the coffers of the (energy-intensive) hi-tech multinational companies.

> Among the customers to be beguiled and taken hostage as soon as possible, there are certainly - due to the volume of financial resources managed, the range of activities to be automated, and the areas of private life over which surveillance practices can be extended and consolidated - governments and public institutions. So it happens, at the same time, that McDonald's realises that artificial intelligence systems are not up to the task, at this moment, of taking an order for a hamburger - and therefore suspends an experiment involving more than 100 points of sale - and that in Italian schools experiments are started that invest such systems with the role of 'new tutors for students'. In order to understand such a picture, it may be useful to ask oneself from which private companies today come the push to introduce immature and non-functional technological products into schools and universities, with what narratives and what reconceptualisation of education companies accompany the promotion of such products, and why schools and universities so easily give in, forgetting the very nature of their business, to the new sellers of 'snake-oil AI', i.e. of applications that do not work (sometimes, simply because they cannot work)[28].

But what can libraries concretely do to make use of artificial intelligence for their activities and to help users make informed use of it? As fact-checkers, librarians are used to critically examining the chain of sources and judging with suspicion information that is lacking. In the latter case, they can teach where and how to look for documents that confirm or disprove the output provided by the machine. Such practices can be implemented both in daily reference activities and through free training courses open to the public, developing responsible search practices. It is important that users and citizens know that they can rely on librarians for these tasks. It is equally important, on the other hand, that librarians are trained in AI Literacy, understood as «teaching how to communicate with the machine, enabling its use, but also in understanding why and on what technological mechanisms, the machine communicates»[29]. Moreover, as librarian Amy Chatfield notes, «there are a lot more context clues behind the scenes than people

[26] See D. Tafani, *Omini di burro. Scuole e università al Paese dei Balocchi dell'IA generativa*, «Bollettino telematico di filosofia politica», 2024, doi: 10.5281/zenodo.13923820.
[27] R. Waelen, *Why AI Ethics Is a Critical Theory*, «Philosophy & Technology», XXXV, n. 9, 2022, doi: 10.1007/s13347-022-00507-5.
[28] D. Tafani, *Omini di burro*, cit.
[29] S. Cuomo, G. Biagini, M. Ranieri, *Artificial Intelligence Literacy, che cos'è e come promuoverla. Dall'analisi della letteratura ad una proposta di Framework* in «Media Education. Studi, ricerche, buone pratiche», XIII, n. 2, 2022, doi: 10.36253/me-13374, pp. 161-172: 163.

might notice, training AI for each specific context and situation might take a lot of time and resources»[30].

On the other hand, librarians and information mediators can benefit for sure from AI-based tools such as Elicit, Perplexity, SciSpace, Scite, Paper Digest, Trinka, Consensus for literature review, research reports, State of the art, comparison, bibliographic research, editing and revision, grammar and plagiarism check. Should we admit that these tools do their job better than librarians? Probably yes, but only as long as we consider the librarian's activity only as source finding and not as a facilitating interaction with other scholars and assisting them for all stages of research. It is therefore not only necessary for librarians to be trained in the potential and limitations of GAI, but also to consider themselves (and be considered by the community) for the contribution they can actually make.

4. Conclusions

Even if Generative AI is undoubtedly capable of facilitating the work of librarians and researchers, this doesn't imply that it is ready to take over the library profession entirely. Instead, evidence suggests that it could be a useful tool to support research activities, cataloguing and so on. Consciously using GAI for information retrieval, translations, and the production of summaries and reports can be very useful to academic research but is not without risk, and many mistakes are still produced by the machine. It's part of librarian skills to be able to identify flaws and red flags in the information source chain. This is why it is important for librarians to be up-to-date on the benefits and biases of these tools, even if they evolve very quickly.

If librarians can benefit from AI, also the reverse is true. Librarians could therefore be good trainers on the biases of GAI as information search tools, and they should be involved as teachers in AI literacy courses. Moreover, librarians cannot be replaced by GAI since their role goes beyond their skills as information professionals, encompassing a human and social dimension that significantly impacts people and communities. Rather than replacing librarians, GAI highlights the need for their expertise more than ever. As guardians of ethical information practices and facilitators of critical engagement, librarians have a key role to play in guiding society through the complexities of the AI age.

[30] M. Hosseini, K. Holmes, *The Evolution of Library Workplaces and Workflows via Generative AI*, cit.

Ethical and aesthetical questions on stock images: the case of AI's depictions[a]

Alberto Romele[*]*, Dario Rodighiero*[†]*, Sabina Rosenbergova*[‡]

Abstract

In questo articolo, gli autori trattano le immagini stock che raffigurano l'IA come un volto o un corpo che subisce un processo di frammentazione in particelle, pixel o voxel. Queste immagini, sostengono, sono i sintomi di una visione del mondo basata sui dati. Nella prima sezione, gli autori discutono le immagini stock dell'IA e rendono conto delle loro analisi qualitative e quantitative di circa 7.500 immagini provenienti dal catalogo online di Shutterstock. Queste analisi hanno evidenziato i volti e i corpi digitalizzati come uno dei temi principali tra le immagini stock dell'IA. Nella seconda parte, gli autori approfondiscono il concetto di digitalizzazione della visione del mondo e offrono alcuni esempi tratti dall'architettura e dal design. Questa seconda sezione include una digressione metodologica, in cui gli autori propongono di articolare l'iconologia di Panofsky e la prospettiva "sintomatica" di Didi-Huberman. In conclusione, gli autori riflettono su un aspetto apparentemente marginale delle immagini stock dell'IA: l'uso abbondante del blu.

Parole Chiave: intelligenza artificiale, blue, iconologia, immagini stock, datificazione globale.

In this article, the authors deal with stock images depicting AI as a face or a body that undergoes a process of fragmentation into particles, pixels, or voxels. These images, they contend, are the symptoms of a datafied worldview. In the first section, the authors discuss stock images of AI and account for their qualitative-quantitative analyses of about 7,500 images from the online catalog of Shutterstock. These analyses have brought out datafied faces and bodies as one of the main themes among stock images of AI. In the second part, the authors elaborate on the notion of datafication of the worldview and offer some examples from architecture and design. This second section includes a methodological detour, in which the authors propose articulating Panofsky's iconology and Didi-Huberman's "symptomatic" perspective. In conclusion, the authors reflect on an apparently marginal aspect of stock images of AI: the abundant use of blue.

Keywords: artificial intelligence, blue, iconology, stock images, worldview datafication.

[a] Received on 13/04/2025 and published on 09/12/2025.
[*] Sorbonne Nouvelle, e-mail: alberto.romele@sorbonne-nouvelle.fr.
[†] University of Groningen, e-mail: d.rodighiero@rug.nl.
[‡] University of Groningen, e-mail: s.rosenbergova@rug.nl.

1. Introduction

Stock images are pre-produced images, available for purchase on agency websites like Getty Images and Shutterstock, which provide financial compensation to the creators upon acquisition. Often dismissed by both public discourse and academic literature as the "wallpaper" of our consumer culture[1], stock images are typically derided for their overtly stereotypical representations of reality. For instance, there are stock images of women inexplicably laughing while eating salad[2] or seemingly unable to drink from a water bottle without spilling it[3]. Nonetheless, stock images pervade our visual landscape: college brochures feature beaming, successful students; magazines are replete with images of busy yet cheerful businesspeople, and so forth. Notably, many of these clichés have facilitated the importation of aspects and practices intrinsic to North American society into other cultures. It could be argued that stock imagery holds nearly as much sway in our lives as Hollywood cinema. Therefore, despite their banality and pronounced kitschiness, stock images warrant our full attention – akin to the scrutiny philosophers like Adorno and Horkheimer applied to the Hollywood studio film system.

It is worth noting that we are not pioneers in recognizing the significance of stock images. Although still somewhat peripheral, literature on the topic is expanding, particularly within the realms of media studies and social semiotics[4]. In this instance, our focus is not on generic stock images but rather on those illustrating Artificial Intelligence (AI). More precisely, we are examining stock images in which AI is depicted as a face or body undergoing a process of composition and decomposition into particles, pixels, or voxels (in other terms, depictions of data), a phenomenon we refer to as well as "face and body fragmentation." We perceive these images as symptomatic of a "datafication" of the worldview, or *Weltanschauung*, as conceptualized by Dilthey and Mannheim. The "datafied" worldview represents a contemporary manner of perceiving, reasoning, acting, and desiring in the world, emerging as an alternative to previously dominant worldviews[5].

Consider the impact of utilizing self-tracking technologies, which has fostered a datafied comprehension of the self, or how the quantity of friends and followers translates into a measurable reputation. For instance, Fourcade and Healy introduced the term *Übercapital* to describe the emergence of a new form of capital that surpasses those

---

[1] P. Frosh, Beyond the Image Bank: Digital Commercial Photography, in M. Lister (edited by), *The Photographic Image in Digital Culture*, Routledge, London-New York 2013, pp. 131-148: 145.

[2] The phenomenon of women laughing alone with salad in stock images has been notably highlighted and satirized in various platforms. For a humorous take on this peculiar trope, see *Women Laughing Alone With Salad* on «The Hairpin», https://www.thehairpin.com/2011/01/women-laughing-alone-with-salad/.

[3] The trope of women seemingly struggling to drink water without spilling it in stock images has also been humorously explored and critiqued. For an entertaining examination of this peculiar stock image trend, see *Women Struggling to Drink Water* on «The Hairpin», https://www.thehairpin.com/2011/11/women-struggling-to-drink-water/.

[4] See, for example, P. Frosh, *The Image Factory: Consumer Culture, Photography, and the Visual Content Industry*, Berg, Oxford 2003; G. Aiello, A. Woodhouse, *When Corporations Come to Define the Visual Politics of Gender: The Case of Getty Images*, in «Journal of Language and Politics», XV, n. 3, 2016, pp. 352-368; C. Thurlow et al., *Visualizing Teens and Technology: A Social Semiotic Analysis of Stock Photography and News Media Imagery*, in «New Media & Society», XXII, n. 3, 2020, pp. 528-549.

[5] A. Romele, *Digital Hermeneutics: Philosophical Investigations in New Media And Technologies*, Routledge, London-New York 2020.

previously identified by Bourdieu, such as economic, cultural, and social capital. They define *Übercapital* as the "long history of a person's recorded actions, built up from traces left on everything from social media to credit bureaus, shopping websites, loyalty programs, courthouses, pharmacies, and the content of emails and chats"[6]. However, the phenomenon under discussion extends beyond merely the datafication of the self; it also encompasses the datafication of others and the environment, resonating with the Heideggerian distinction between *Selbswelt*, *Mitwelt*, and *Umwelt*. It is not only technological but also artistic, cultural, and epistemological – consider, for example, the rise of data-driven research[7]. This article will delve into the aesthetics of such a phenomenon.

The article is structured into two main sections. The initial section delves into stock images of AI, presenting our qualitative-quantitative analyses. These investigations highlight the fragmentation of faces and bodies as a predominant theme in AI stock images. Additionally, this section offers insights into the visual representability of AI. The subsequent section interprets these images as indicative of a transformative shift, specifically, the rise of a datafied worldview. Within this section, we discuss our methodological approach, drawing inspiration from Erwin Panofsky's iconology and Georges Didi-Huberman's critique of Panofsky's framework. In our concluding remarks – which we believe transcend a traditional conclusion – we engage with Didi-Huberma 's reflections on Fra Angelico, focusing on a nuanced aspect of these images: the prevalent use of the color blue.

2. Facial and Body Datafication

Disciplines such as science and technology studies have long demonstrated an interest in images for two compelling reasons. Firstly, emphasizing the role of images inherently underscores the role of technology in the development of scientific knowledge. Succinctly, exploring images illustrates how science has become progressively and persistently reliant on technical instruments, including those technologies that generate images of objects and phenomena that might otherwise remain elusive. Secondly, this focus aligns with the imperative to transcend certain excesses of the logocentrism and textocentrism that permeated much of the human and social sciences in the twentieth century.

However, despite a heated interest in imagery on the part of these disciplines, it is noteworthy that there is also a great void here. On the one hand, scholars have been interested in images produced by scientists for other scientists via technical instrumentation[8]. On the other, they have been interested in producing scientific images by artists, mostly in collaboration with scientists[9]. In short, it seems that the attention of researchers goes towards those images of science and technology under the aegis of two regimes of truth: the regime of scientific reference and the regime of aesthetic taste. It does not matter if the reference is accepted in a naïve way or if it is criticized; it does not matter if the aesthetic taste, when applied to science and technology images, can still be

[6] M. Fourcade, K. Healy, *Seeing Like a Market*, in «Socio-Economic Review», XV, n. 1, 2017, pp. 9-29: 18.
[7] S. Leonelli, *Scientific Research and Big Data*, in «Stanford Encyclopedia of Philosophy», Stanford University, Stanford 2020.
[8] M.E. Lynch, S. Woolgar (eds.), *Representation in Scientific Practice*, MIT Press, Cambridge (MA)-London 1990; C. Coopmans et al. (eds.), *Representation in Scientific Practice Revisited*, MIT Press, Cambridge (MA)-London 2014.
[9] P. Galison, C.A. Jones (eds.), *Picturing Science, Producing Art*, Routledge, London-New York 1998.

disinterested (as the Kantian tradition would like) or not. In both cases, discussion and interest remain within these two regimes of truth. In this way, navigating between the Scylla of technical images and the Charybdis of artistic images, we overlook a category of science and technology images that are neither produced by scientists nor boast artistic claims. These images, more artisanal than artistic, find their pinnacle in today's proliferation of stock images related to science and technology.

In this context, our discussion centers around AI-related stock images. However, before delving into our primary subject, we wish to briefly explore the concept of AI's representability. Presently, when we refer to AI, we predominantly allude to machine learning algorithms. We posit that there are three potential methods to describe today's AI: 1) through the algorithm, which can be embedded in various forms in the computer code. This approach, however, is unsatisfactory for two main reasons: it is not comprehensible to non-experts, and representing the algorithm does not equate to representing AI – akin to equating the representation of the brain with intelligence; 2) through the technologies in which AI is embedded, such as drones, autonomous vehicles, and humanoid robots. However, depicting AI through these technologies does not genuinely represent it, as nothing indicates that this technology is AI-driven rather than merely an empty shell; 3) by foregoing the representation of the "thing itself" and focusing instead on expectations or imaginaries. The majority of stock images and other popular AI representations fall into this category. This threefold distinction is idealtypical (in the vein of Weber's ideal types) as these three levels often intermingle. For instance, lines of code are colored, drones fly over verdant meadows, robots exhibit a flawless white surface, and even the most abstract images subtly allude to what exists or is being developed – a touch screen, a neural network, and so forth.

Currently, researchers tend to evaluate images of AI – and technology in general – based on their ability to represent the "thing itself" from ontological, ethical, and aesthetic perspectives. Consequently, there is a preference for the first method of representation over the second, and the second over the third. An image is deemed more "true", "good," and "aesthetically appreciable" the closer it is, and thus, the more faithful it is, to the entity it intends to represent – a phenomenon we term "referentialist bias." However, given the aforementioned considerations, referentialism proves ineffective in the context of AI images, as none can accurately or faithfully approach AI. Our proposition is not to condemn AI images but to redeem them by eschewing referentialism. Conversely, if an aesthetics, which encompasses ethics and ontology, of AI images exists, its objective is not to portray AI technology per se. Rather, its aim, a concept we will elaborate upon in the conclusion, is to "give rise to thought".

Enter "Artificial Intelligence" into any search engine and select the "Images" option. This action will yield a variety of images: half-flesh, half-circuit brains; humanoid robots interacting with touch screens; lines of code wafting through space; and variations of Michelangelo's *The Creation of Adam*, reimagined with human-robot interactions.

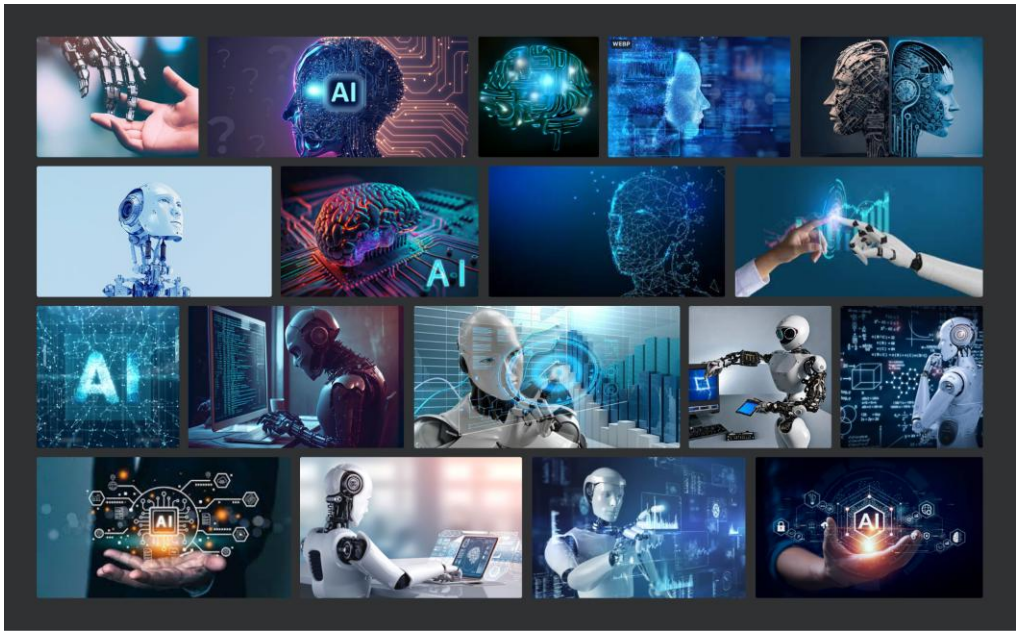

Figure 1: A visual montage showcasing the top image results from a Google search using the keyword "Artificial Intelligence." The images collectively illustrate various conceptualizations and symbolic representations of AI, highlighting both technological and humanistic elements.

Stock images depicting AI are widely used not only in popular contexts but also in science communication contexts by research organizations and institutions. Particularly striking are instances where the ethics of AI is discussed, yet the ethics of visual science communication on AI is seemingly disregarded[10].

The sheer volume of AI-related stock images is staggering; for instance, a July 2025 search for "artificial intelligence" on Shutterstock yields 435,672 results. To navigate beyond a mere qualitative analysis of a modest subset of these images in our research, we employed automated methods. Initially, we utilized the web crawler ShutterScrape (Lin), enabling the massive download of images and videos from the American provider, thereby acquiring approximately 7,500 AI-related stock images. Subsequently, we employed PixPlot (Duhaime), a tool developed within the Digital Humanities Lab (Yale University), to visualize extensive image collections within an interactive WebGL scene. Each image undergoes processing with an Inception Convolutional Neural Network, trained on ImageNet 2012, and is projected into a two-dimensional manifold using the UMAP algorithm of dimensionality reduction (McInnes), ensuring similar images are proximate to each other. The resultant visualization can be seen below and accessed via the link in the references to zoom into detailed areas[11].

[10] Examples of this discrepancy in ethical consideration can be seen in various images: the cover of the Oxford Handbook of Ethics of AI, https://www.instagram.com/p/CPH1wmr216/, the now-obsolete EU page featuring the Ethical Guidelines for Trustworthy AI, https://www.instagram.com/p/CPH8xoCLTm7/, and the site of the SHERPA project, an "EU-funded project analyzing how AI and big data analytics impact ethics and human rights," https://www.instagram.com/p/CVA9Y-dIoRv/.
[11] A. Romele, *Digital Hermeneutics*, cit.

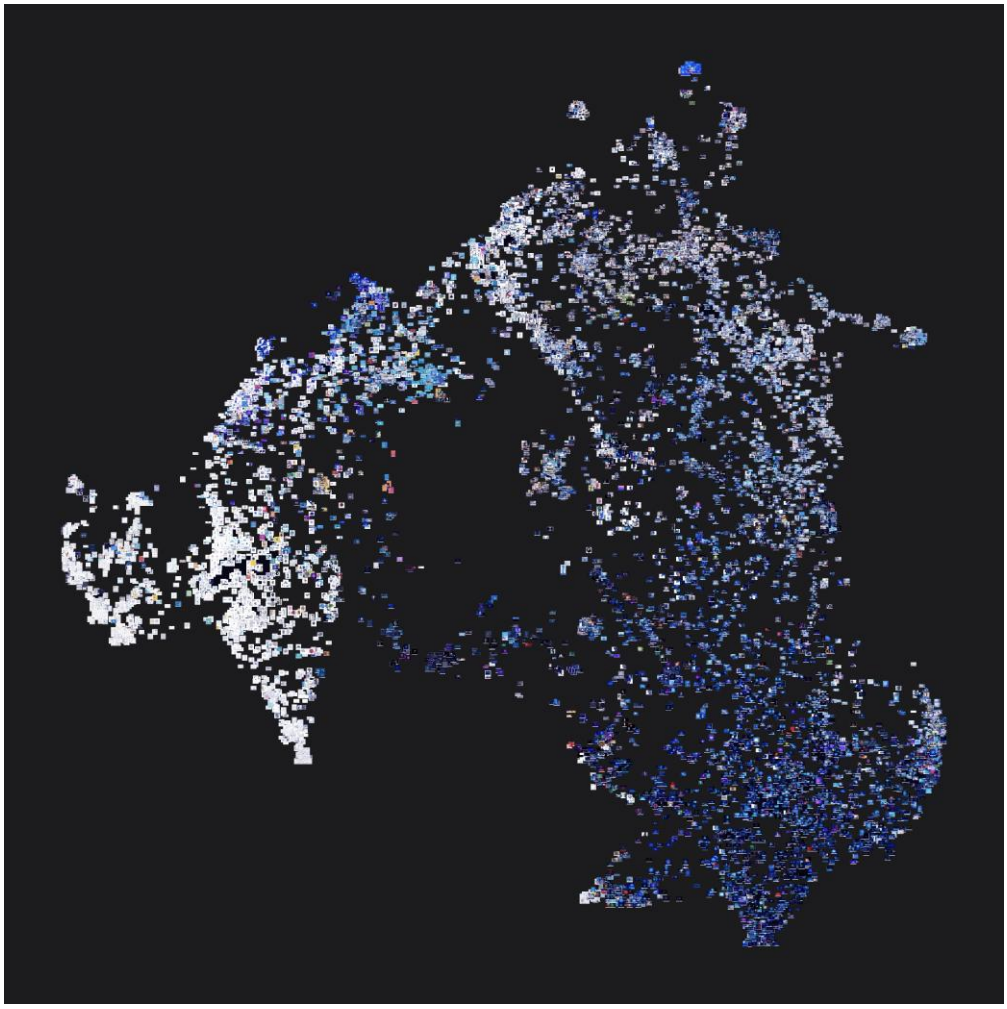

Figure 2: Exploring the visual landscape of AI through PixPlot. This image provides a glimpse into our curated catalog. Within the interactive PixPlot interface, each cluster of images, organized through machine learning algorithms, can be closely examined by users, enabling a deeper dive into the visual themes and variations present within the extensive collection of AI-related stock images.

The AI stock images were systematically divided into ten clusters, each manually labeled as follows: 1) background, 2) robots, 3) brains, 4) faces and profiles, 5) labs and cities, 6) line art/line drawing, 7) Illustrator, 8) people, 9) fragments, and 10) diagrams. PixPlot neither suggests automatic labeling for clusters nor provides explanations for the specific groupings of images. Consequently, the labeling work was deductively executed manually, guided by insights collected through careful observation. Interestingly, an AI algorithm undertakes the task of recognizing and categorizing images related to artificial intelligence. This technology increasingly self-selects its visual representations: stock images are procured through search engines on stock imagery agency websites. The most purchased images appear as first-page results, adhering to the algorithmic logic of Google's PageRank. This logic compels image producers to create visuals with the potential for success, elucidating the recurrence of specific themes despite the abundance of images[12]. Notably, PixPlot discerns similarities between images by reducing them to a pixel vector, serving as a quintessential example of the datafication process under scrutiny in this article.

[12] *Ibidem*.

Images exist not as *qua* images but as aggregates of pixels. Digital images are data, and all digital imaging, whether consciously or not, constitutes an act of data processing; the visual surface of digital images reveals nothing about their nature. Indeed, «[digital] images are far more closely related to spreadsheets and statistical formulas than to photographs» (May 50).

Quantitatively speaking, the "fragmentation" was one of the clusters automatically identified by PixPlot. This is particularly noteworthy given that existing literature and critical discussions on AI's visual representations have predominantly centered on different types of images, especially those depicting AI through humanoid robots, chatbots, and virtual assistants[13]. Turning our attention to this cluster composed by fragments, we delve into these visual "dispersions" by selecting three representative images illustrated below. These images exude ambiguity, concurrently evoking two notions: the emergence of a machine transitioning into a quasi-conscious, quasi-human state, and the fragmentation of the human into a quasi-digital machine. A face or a body, often androgynous or female, materializes, constructed from particles, pixels, or voxels. The gaze, often averted, usually to the right, implies a future unfolding, while in instances where the gaze meets the viewer, as in one of the images linked in the note, the tone shifts to something more ominous, presenting a machine that directly challenges the observer. The colors tend to be sterile, oscillating between white and blue, a common palette for stock images of artificial intelligence and other emergent technologies like cloud and quantum computing. These visuals subtly allude to a transition, a metamorphosis from human to machine, and vice versa.[14]

Figure 3: This triad of images, extracted from our cluster focused on fragmentation, showcases various visual interpretations of the transition between human and machine. Each image, while embodying a distinct aesthetic, subtly navigates through the concept of fragmentation and defragmentation, offering a visual exploration into the datafied worldview where humans, machines, and their respective realities intertwine.

Our thesis posits that images illustrating the transition from human to machine – and conversely – from machine to human, through processes of datafication (fragmentation or defragmentation), imply that these fragments or data form the commonplace reality of both humans and machines and, ultimately, of the world itself. This perspective also substantiates why the cluster of fragments encompasses more than

[13] S. Cave, K. Dihal, *The Whiteness of AI*, in «Philosophy & Technology», XXXIII, n. 4, 2020, pp. 685-703.
[14] Visual social semiotics offers a set of tools for analyzing the internal dynamics of images and their effects. The relationship between gazes is only one among them. The use of colors is another. Added to this are vectors and directions, framing, exploitation of depth, etc. See G. Kress, T. van Leeuwen, *Reading Images: The Grammar of Visual Design*, 3rd edition, Routledge, London-New York 2020.

just faces and bodies. In their stark simplicity and kitschness, these stock images become vessels for a world picture or worldview, which we specifically term a "data-driven" or "datafied" worldview.

3. Datafication of the world

A methodological *détour* is requisite before delving into the discussion on the datafication of the world. The thesis of the "data-driven" or "datafied" worldview draws inspiration from Erwin Panofsky's work in iconology and his theses on *habitus* or "mental habit" from the twelfth-thirteenth century. In the introduction to one of his books, Panofsky utilizes the well-known example of a man lifting his hat to greet an acquaintance[15]. This example serves to delineate three levels of observation and interpretation of a work of art: 1) a perceptual level, where one identifies a simple series of colors, lines, and shapes in the gesture; 2) a social level, where one recognizes the gesture as a greeting, necessitating familiarity with the practical world of objects and events, and the "more-than-practical" world of customs and cultural traditions characteristic of a specific civilization; and 3) an "intrinsic" or "content" level that transitions from iconography to iconology, where both specific elements (how exactly did the man raise his hat?) and more general ones are considered. At the third level, art history reaches its terminus, aiming to perceive in a single work of art the style and habit of a time, and the underlying principles that dictate its existence.

Panofsky's iconology transcends the boundaries of art history, extending to all cultural expressions of an epoch. For instance, in Gothic Architecture and Scholasticism[16], he posits a hypothesis that, during the twelfth and thirteenth centuries, the connection between Gothic art and Scholastic philosophy was more tangible than mere parallelism, yet also more general than a direct influence that scholastic thinkers might have exerted on artists and craftsmen. Between scholastic intellectuals and artists, Panofsky argues, there existed not a cause-and-effect relationship (which might suggest, for example, that architects were avid readers of scholastic treatises) but one of diffusion. This diffusion is precisely what Panofsky terms *habitus* or "mental habit". The mental *habitus* shared by Gothic architects and scholastic philosophers, as well as their respective works, is grounded in a rejuvenated trust in reason, perceived as capable of substantiating anything that can be deduced from principles distinct from faith. Specifically, Panofsky identifies three elements or principles of similarity between Gothic architecture and scholastic texts: 1) totality or sufficient numbering, 2) arrangement according to a homologous system of parts and subparts, and 3) distinctiveness and deductive power. Exemplary is the analogy between the micro-architectonic division into logical and determined levels at Portal of the Last Judgment of Notre Dame Cathedral in Paris and the text structure – the uniform division and subdivision of the logical sections – of Thomas Aquinas's Summa.

A potent critique of the Panofskian method, particularly in relation to art and images more broadly, was formulated by Didi-Huberman. He posits that the predominant history of art, which was indebted mainly to Panofsky's method developed after his refuge from Germany for the United States, is largely neo-Kantian in its inspiration. Didi-Huberman scrutinizes the two versions of Panofsky's work, one from 1932 in German and the other,

[15] E. Panofsky, *Studies in Iconology: Humanistic Themes in the Art of Renaissance*, Westview Press, Boulder 1972.
[16] Id., *Gothic Architecture and Scholasticism*, Archabbey Publications, Latrobe 2005.

an English article from 1939, in which the example of the gentleman raising his hat is introduced – an example sourced from Mannheim.[17] Didi-Huberman notes the presence of terms in the 1932 version, influenced also by Mannheim, such as "supreme region" and "sense of essence" to denote the ultimate objective of art history. However, the approach to realizing this synthesis project was markedly different in 1932. In Didi-Huberman's words: "This project was radical, it was different: uneasy, traversed by a force that, far from being pedagogical, was questioning, almost convulsive [...] and quite authentically philosophical"[18]. Transitioning from Germany to the United States, what perishes from Panofsky's method is the antithesis, supplanted by an optimistic, positive, and even positivistic synthesis.

Panofsky was part of an intellectual movement that sought to historicize, socialize, and culturalize Kant's schematism in 20th-century Germany, reminiscent of the works of Ernst Cassirer, and similarly in France, akin to the approaches of Émile Durkheim and Marcel Mauss. Didi-Huberman posits that the efforts of Panofsky, as well as those of Cassirer, Durkheim, Mauss, and extending to Bourdieu, to historicize, socialize, and culturalize the Kantian schematism were insufficient to transcend the synthetic temptation intrinsic to Kant's thought[19]. While Didi-Huberman acknowledges the distinction between Panofsky's historicism and Kantian apriorism and a psychologizing interpretation of Kant, he contends that Panofsky merely substituted one form of universalism (the transcendental or psychologizing kind) with another, rooted in historical context. Panofsky's iconology, in Didi-Huberman's view, is as much a transcendental synthesis as Kant's transcendentalism. In a footnote, while discussing Panofsky's renowned interpretation of Titian's *Allegory of Prudence*, Didi-Huberman notes, "he was (they were) looking not at the painting itself – with its dark, evenly colored focal mass – but rather at a black and white photograph of it"[20]. Furthermore, he cites a passage from Panofsky's 1932 article, which states, "the greatness of an artistic production is ultimately dependent upon the quantity of "*Weltanschauung*-energy" that is incorporated into the worked material and radiates back from it to the spectator"[21].

If Didi-Huberman's critique holds merit, then any endeavor to extrapolate an entire worldview from a handful of images, such as those discussed in the previous section, would amount to a synthetic act as problematic as Panofsky's. This would be further exacerbated by the lack of nuance compared to Panofsky's analyses. However, a solution to this methodological dilemma is offered by Didi-Huberman himself. In the concluding section of his book, he emphasizes the role of "the negative" in imagery, advocating for a focus on the symptom – or lapsus – over the sign and the affirmative narrative that a work of art presents. Didi-Huberman argues for concentrating not on the synthesis but on the elements that render such synthesis incomplete, and ultimately, unattainable. While he

[17] The 1939 version, which underwent subsequent transformations in 1955 and 1962, is featured in the introduction to *Studies in Iconology* (see E. Panofsky, *Studies in Iconology*, cit.). The 1932 version was initially published in the journal «Logos» (XXI) with the title *Zum Problem der Beschreibung und Inhaltsdeutung von Werken der bildenden Kunst*, and was later incorporated into *Perspective as Symbolic Form*, see E. Panofsky, *Perspective as Symbolic Form*, Zone Books, New York 1997.

[18] G. Didi-Huberman, *Confronting Images: Questioning the Ends of a Certain History of Art*, Pennsylvania State University Press, University Park 2005, p. 98.

[19] Ivi, p. 168.

[20] Ivi, p. 293.

[21] Ivi, p. 126.

acknowledges Panofsky's pioneering role in introducing the concept of the symptom at the level of iconology, he criticizes him for immediately closing off this new avenue. According to Didi-Huberman, Panofsky did not treat symptoms as irreducible entities but rather reduced them to documents of a "homogeneous" worldview.

That said, however, we believe that Didi-Huberman's perspective is not as far removed from Panofsky's, or from others who have engaged in the historicization, socialization, and culturalization of Kantian transcendentalism, as he might suggest. The key difference between the two lies in their respective emphases: Didi-Huberman focuses on the impossibility of synthesis (though he synthesizes by making symptoms and lapsus foundational elements of a method, much like Freud), while Panofsky, along with Cassirer, Durkheim, and Bourdieu, emphasizes the potential for synthesis, albeit one that is always constrained by cultural, temporal, and social factors.

Instead, our methodological approach aims to navigate between Panofsky and Didi-Huberman, employing the concept of the symptom – as well as the notions of trace and remainder (or "*reste*" in French) – to advocate for a "fragile epistemology." This would be a synthesis that is neither absolute nor unattainable, but rather perpetually open-ended. In this vein, our approach aligns with the evidential paradigm[22], which emphasizes the importance of traces and clues. While we use a select number of images to illustrate a broader worldview, we remain open to the inclusion of other images and cultural productions, as well as alternative interpretations of these traces or clues. To elucidate our concept of a "data-driven" or "datafied" worldview, we turn to the book *The Second Digital Turn*[23]. His approach is particularly compelling because it echoes Panofsky's method to some extent, drawing its validation from a cross-disciplinary comparison that includes fields like industrial design, architecture, and philosophy.[24] Mario Carpo argues that the essence of this second digital turn can be encapsulated in the phrase "Search, Don't Sort." This was the tagline beneath Gmail's logo when the service debuted in 2004, emphasizing its revolutionary feature: the ability to search the full text of all email messages, whether sent or received, by words or numbers. Carpo suggests that this automated full-text search capability on a corpus of unsorted data is a more effective retrieval tool than the traditional method of first categorizing items by topic and then searching within those categories. In essence, given the vast amount of data available to us, it is more efficient to entrust the task of finding a personalized, temporary, and context-specific order and meaning to a machine or algorithm. Unlike the tree-like structures designed for human cognition and memory, computers operate differently.

Carpo contends that this emerging preference for rawness and disorder is evident in modern design and architecture. He points to the spline curve, a feature of early CAD/CAM technology, as an exemplary manifestation of the "small data logic" – which could be described as both informational and linear – that characterized the digital style of the early 1990s. Carpo also acknowledges the influence of Deleuze's concept of the "fold"

[22] C. Ginzburg, Clues: Roots of an Evidential Paradigm, in *Clues, Myths, and the Historical Method*, Johns Hopkins University Press, Baltimore 1989, pp. 96-125.

[23] M. Carpo, *The Second Digital Turn*, MIT Press, Cambridge MA and London 2017.

[24] However, a key distinction exists. While Carpo often interprets changes in architecture and design as direct outcomes of specific technical advancements – particularly digital ones – in the practices of architects and designers, his argument can sometimes come across as deterministic. We lean more towards a Panofskian approach, which focuses on identifying "family resemblances" among various elements without necessarily attributing them to a cause-and-effect relationship.

in shaping this curvilinear aesthetic. Deleuze's fold is a mathematical curve that he likened to continuous functions and Leibniz's development of differential calculus. In a similar vein, spline modelers are capable of converting any arbitrary set of points into seamlessly curved lines, making a spline the smoothest possible line that can connect multiple fixed points. These spline modelers found their way into architecture through the work of Pierre Bézier and Paul de Casteljau for Renault and Citroën, and subsequently into the designs of Frank Gehry and his Los Angeles studio. Iconic works like the Fish sculpture in Barcelona and the Guggenheim Museum in Bilbao stand as testaments to this early digital-driven architectural style. According to Carpo, this kind of

> Small data-oriented architecture" and design has already been overcome by the use of big data and algorithms of artificial intelligence to engage the messy discreteness of nature as it is, in its pristine, raw state – without the mediation or the shortcut of elegant, streamlined mathematical notations. The messy point clouds and volumetric units of design and calculation that result from these processes are today increasingly shown in their apparently disjointed and fragmentary state; and the style resulting from this mode of composition is often called voxelization, or voxelation[25].

Carpo cites a range of examples to illustrate this trend, including the Computational Chairs Design studies by Philippe Morel of EZCT Architecture & Design, which were created using genetic algorithms. First showcased at the 2013 ArchiLab exhibition in Orléans, France, these studies were among the earliest expressions of this new design philosophy. This approach has since been adopted by a diverse group of architects, designers, and artists, including Alisa Andrasek and Jose Sanchez, Marcos Cruz and Marjan Colletti, Andrew Kudless, David Ruy and Karel Klein, as well as Jenny Sabin and Daniel Widrig.

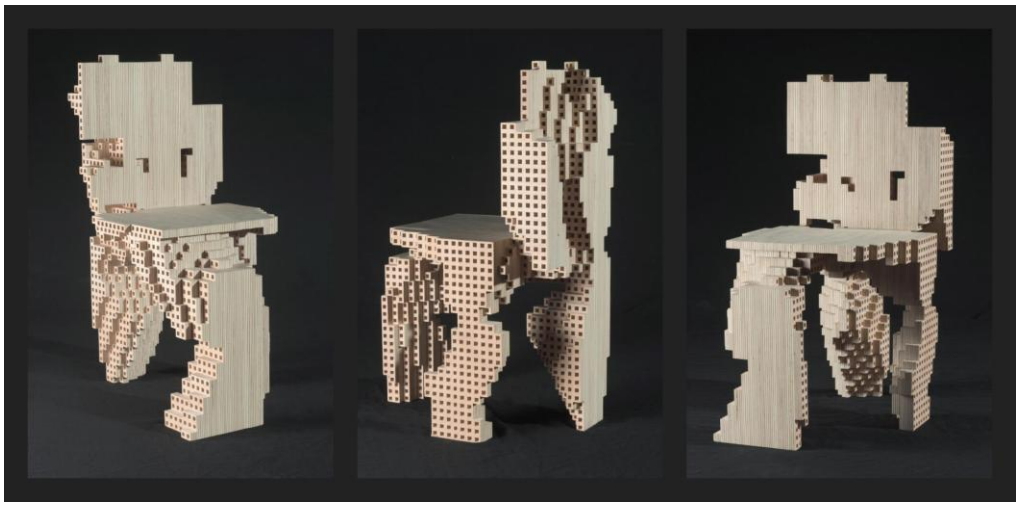

Figure 4: "Modèle test multi chargement 'Bolivar—320 après 320 générations (32,000 évaluations structurelles)" by EZCT Architecture & Design Research. This prototype is part of the series "Computational Chair Design Using Genetic Algorithms".

[25] Ivi, ch. 2.7.

We can draw an analogy to Panofsky's three-step method of visual interpretation by suggesting that a "data-driven" or "datafied" worldview also rests on three foundational elements: 1) a principle of emergence, positing that meaning arises from an emergent, partial order rather than a rigid structure that excludes exceptions; 2) a principle of instability, which acknowledges that the relationships between data points are in constant flux due to the ongoing influx of new data and evolving algorithms; and 3) a principle of deduction that operates without a clear understanding of underlying causes or rules. For instance, consider machine learning algorithms like PixPlot, which reduce high-dimensional data to a two-dimensional Cartesian plane. In this process, humans play no active epistemological role, which accounts for the difficulty in comprehending the mechanics of the operation.

4. More than a conclusion: nel blu dipinto di blu

In conclusion, we would like to return to the specific issue of stock images of AI. It is a common idea that we must have "more accurate" images of AI. We hypothesize that today, concerning AI, we are very much in need of "pensive" images. The term is taken from Jacques Rancière[26], according to whom an image is pensive to the extent that it can bring together different regimes of expression without ever synthesizing them. In other words, an image is pensive insofar as it forms metaphors that are always open and never exhausted according to different spatial and temporal planes. For example, rephrasing Didi-Huberman, AI stock images do not give rise to thought as their meanings are so simplistically precise that there is no margin left for a remainder, trace, or symptom. The display is absolute, and nothing is denied to viewers; however, in this way, nothing is given to viewers to think about. In other words, the display distinguishes between the visual regimes of erotism and pornography – for a deeper discussion on this topic in a different context[27]. The visual regime of erotism can be summarized by the formula "to know how (or be able) not to look at." The visual regime of pornography can be resumed instead in the formula "to not know how (or not be able to) not to look at." It means that while erotism is based on the persistence of a gap between the visible and the invisible, pornography is the apotheosis of visibility that annihilates invisibility. We argue that stock images of AI are pornography of AI depiction.

The visual allure of AI stock imagery is evident, for instance, in the fact the color blue is far the most frequently used color among these images. This evident preference deserves further consideration. Here, again, Didi-Huberman's opening pages of *Confronting Images* can be inspirative of how we can frame this question. In these, the philosopher discusses Fra Angelico's *Annunciation* and the surprising omnipresence of white color that is diffused throughout the space of the fresco. He states that a Panofskian iconological interpretation of this artwork falls short as his narrative is conveyed in a sparse and unembellished manner. Fra Angelico appears ill-suited to capture the essence of fifteenth-century Italian painting, which is known for its rich variety – encompassing everything from apocryphal details and illusionist elements to complex spatial configurations and everyday objects. Didi-Huberman contends that the alternative "is based on the general hypothesis that the efficacy of these images is not due solely to the transmission of

[26] J. Rancière, *The Emancipated Spectator*, Verso, London-New York 2008.
[27] See M. Leone, *Annunciazioni: Percorsi di semiotica della religione*, Aracne, Roma 2014, pp. 603-604.

knowledge – visible, legible, or invisible – but that, on the contrary, their efficacy constantly operates in the intertwining, even the imbroglio, of transmitted and dismantled bits of knowledge, of produced and transformed not-knowledges"[28].

Figure 5: Fra Angelico's *Annunciation*, painted in 1440 in the Dominican convent of San Marco in Florence. The painting serves as a focal point for discussing the role of color, particularly white, in conveying complex layers of meaning and emotion. Unlike the anesthetizing blue prevalent in stock AI imagery, the white in this fresco opens up multiple avenues for interpretation and contemplation (Wikimedia).

On the question of the white color in the fresco, which extends throughout the cell where the Archangel meets the Virgin Mary, the French philosopher and art historian asks: "what to make of this white?". And he argues that this white is not a void or lack (*manque* in French); rather, it is substantive: "It is not visible in the sense of an object that is displayed or outlined; but neither is it invisible, for it strikes our eye, and even does much

[28] G. Didi-Huberman, *Confronting Images*, cit., p. 16.

more than that. It is material. It is a stream of luminous particles in one case and a powder of chalky particles in the other. It is an essential and massive component of the work's pictorial presentation. Let us say that it is visual"[29]. Didi-Huberman unhesitatingly labels this white a *symptom*, defining it as "the suddenly manifested knot of an arborescence of associations or conflicting meanings"[30]. Finally, the white in Fra Angelico's work is far from representing a void; rather, it catalyzes contemplation. Specifically, the white in the fresco speaks to the theological interpretation of the *Annunciation*, which the Dominican Albert the Great and those who followed his spiritual tradition as Fra Angelico, viewed not merely as a singular event, but as an "absolutely extravagant efflorescence of inclusive or associated meanings, of virtual connections, of memories, of prophecies [...]"[31].

As previously noted, the problematic nature of these AI images becomes evident in the pervasive use of blue in stock pictures. Unlike the color white in Fra Angelico's work, which invites contemplation and opens up multiple avenues of interpretation, the use of blue tends to limit and confine meaning. This is akin to what sociologists of science and technology refer to as "lock-in." However, in this context, "lock-in" is not merely a technical, economic, or material concept; it extends to the realm of imagination and symbolism. When these symbolic forms are embedded in cultural expressions, they significantly influence both societal expectations and the trajectory of innovation.

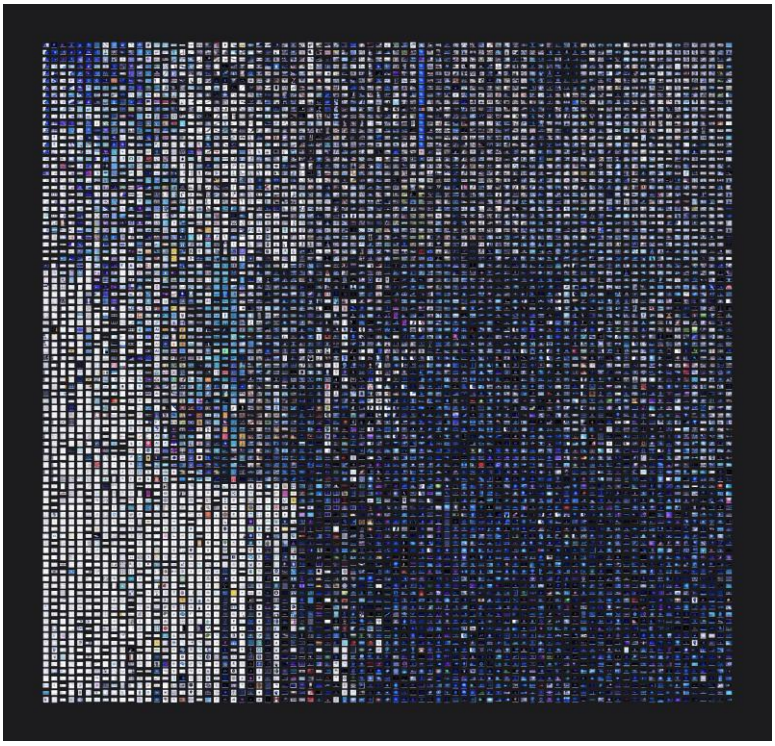

Figure 6: Dominance of blue in AI stock images, contrasting with the more nuanced use of white in artistic works like those of Fra Angelico[32].

[29] Ivi, p. 17.
[30] Ivi, p. 19.
[31] Ivi, p. 22.
[32] For further discussion on the use of white in AI visual representations, see S. Cave, K. Dihal, *The Whiteness of AI*, cit.

Significantly, in the cultural context of Europe, the blue color is of the utmost importance. Its prominent position, often intertwined with spiritual dimension, is clear when looking at the artistic production from Antiquity, through its usage in the sacred spaces of apsidal mosaics of early medieval churches, its prominence in the stained glass windows of the Gothic cathedrals and its migration on the garment of Virgin Mary becoming its topical color as well as its association with the kings of France. In the secularized world, then, the popularization of blue starts with *Young Werther* and *Madame Bovary* and ends with the Levi's blue jeans industry and IBM, referred to as the Big Blue. To this day, blue is the statistically preferred color in the world. According to the French historian Pastoureau's book *Blue: The History of a Color*, the success of blue is not the expression of some impulse, as could be the case with red; instead, one gets the impression that blue is loved because it is peaceful, calming, and anesthetizing. It is no coincidence that blue is the color used by supranational institutions such as the United Nations, UNESCO, and the European Community. In Italy, the police force is blue, which is why policemen are disdainfully called "Smurfs".

Considering the anesthetizing effect of blue and its overabundance in AI, we can argue that the problem with stock AI images is that, instead of provoking debate and "disagreement," they lead the viewer into forms of acceptance and resignation. Rather than equating experts and non-experts, encouraging the latter to influence innovation processes with their opinions, they are "screen images" – following the etymology of the word "screen", which means "to cover, cut, and separate". The notion of "disagreement" or "dissensus" (*mésentente* in French) is taken from Jacques Rancière (*Disagreement*), according to whom disagreement is much more radical than simple "misunderstanding (*malentendu*)" or "lack of knowledge (*méconnaissance*)". As the words indicate, these latter notions are failures of mutual understanding that knowledge can overcome if treated correctly. Interestingly, much of the literature interprets science communication precisely to overcome misunderstanding and lack of knowledge. Instead, we propose an agonistic model of science communication, particularly in the use of images. These images should not calm down but flourish in an agonistic conflict (i.e., a conflict that acknowledges the validity of the opposing positions but does not want to find a definitive and peaceful solution to the conflict itself) – this point has been fully developed by Romele[33].

Stock images of AI are "anaesthetical," a term combining "aesthetics" and "anesthetics." By "anaesthetical," we mean that AI stock images do not promote primary forms of participation but rather "put them to sleep." Just as Fra Angelico's white expanded throughout the fresco and, beyond the fresco, it is possible to think that the anesthetizing effects of blue expand to the subjects and the entire media communication environment in which these AI images proliferate.

Moving from the figure of the images (the face and body fragmentations) to an element of the background (the blue), we are making a gesture similar to Didi-Huberman[34], who looks at some red blotches in the painted garden in his analysis of Fra Angelico's fresco *Noli me tangere*. According to Didi-Huberman, these red blotches are not simple flowers because they are drawn in the same way Fra Angelico draws the stigmata on Christ's feet and hands. The red blotches are an extension of Christ's stigmata: "Christ is here represented in the emblematic act of 'sowing' his stigmata in the garden of the earthly

[33] A. Romele, *Digital Hermeneutics*, cit.

[34] G. Didi-Huberman, *Fra Angelico: Dissemblance and Figuration*, University of Chicago Press, Chicago 2009.

world, just before going to rejoin the right hand of his father in Heaven"[35]. For Didi-Huberman, the background contains the real message of Fra Angelico's painting, which is none other than the mystery of the incarnation – and this also justifies the white used in the painting we discussed above.

As far as we are concerned, however, we do not want to absolutize the background. Instead, we contend that the background grants a new strength to the figure – again, it is a sort of equilibrium between Panofsky and Didi-Huberman. So those stock images, which represent AI and its relation to us and the world in terms of fragmentation, end up seeming obvious: the worldview's datafication appears ineluctable. In conclusion, we argue that stock images of AI do not limit themselves to embedding a datafied worldview from elsewhere – e.g., from the minds of Google's and Meta's engineers. Through their anesthetics, stock images actively contribute to its development and success.

[35] Ivi, pp. 20-21.

Didattica e intelligenza artificiale: risvolti etici, problemi di privacy e sorveglianza, manipolazione dei dati[a]

Patrizia Natale[*]

Abstract

L'introduzione dell'intelligenza artificiale (IA) nella didattica sta rivoluzionando l'educazione, offrendo personalizzazione e ottimizzazione gestionale, ma sollevando importanti questioni etiche. Tra queste emergono i rischi legati a privacy, sorveglianza, manipolazione e "bias" algoritmici, che possono generare discriminazioni razziali, di genere, economiche e linguistiche, ampliando le disuguaglianze preesistenti. Inoltre, la gestione dei diversi dati degli studenti, spesso minorenni, raccolti dai programmi di IA richiede trasparenza e sicurezza, nel rispetto di normative come il GDPR e l'AI Act. Per evitare che l'IA rafforzi divari sociali e stereotipi, è fondamentale una progettazione etica, il monitoraggio continuo degli algoritmi e la collaborazione tra scuole, aziende e autorità. Serve integrare nei curricula l'educazione digitale e la consapevolezza critica sull'uso della tecnologia, ricordando che l'IA deve restare uno strumento al servizio dell'apprendimento e non un fine. Solo con una governance solida e condivisa si potrà garantire un'istruzione equa, inclusiva e rispettosa dei diritti di tutti.

Parole Chiave: didattica ed intelligenza artificiale; bias; disuguaglianza algoritmica; protezione dei diritti individuali; intelligenza artificiale; controllo degli algoritmi.

This paper explores the limitations of algorithmic fairness, particularly the "impossibility theorem of fairness", and discusses how a structural understanding of justice can address the related ethical concerns. After presenting the main models of algorithmic fairness, I argue that they overlook key justice concerns by prioritizing outcome-based metrics and isolating decision-making from broader socio-historical contexts. Furthermore, when base rates differ, it becomes impossible to satisfy more than one fairness metric simultaneously. To address these shortcomings, I propose integrating algorithmic fairness with Iris M. Young's notion of structural injustice, which accounts for entrenched inequalities rooted in the interplay of behaviors, norms, and institutions. This approach situates algorithms within their broader socio-historical context, emphasizing systemic factors that influence decision-making and perpetuate unjust outcomes. I further contend that a structural perspective assigns algorithms a twofold role, particularly in contentious cases where ethical controversies are at play. First, a diagnostic function: by exposing underlying ethical imbalances and biases, algorithms can highlight critical areas for systemic reforms. Second,

[a] Saggio ricevuto in data 10/03/2025 e pubblicato in data 09/12/2025.
[*] Docente di informatica presso istituto di istruzione superiore, e-mail: pattnatale@gmail.com

they can serve as evaluative tools, enabling the assessment and prioritization of fairness metrics on a case-by-case basis.

Keywords: education and artificial intelligence, bias, algorithmic inequality, protection of individual rights; artificial intelligence, algorithmic control.

1. Introduzione

Le metodologie didattiche stanno subendo profonde trasformazioni con l'introduzione dell'intelligenza artificiale (AI), e continuamente vengono evidenziati i benefici derivanti dall'utilizzo dell'AI per giungere ad una vera individualizzazione dell'educazione rispettando ed esaltando le competenze e conoscenze pregresse, il background culturale le esperienze vissute dai discenti. L'uso dell'AI, infatti, permette la personalizzazione dell'apprendimento e il miglioramento dei processi di gestione del gruppo classe elaborando percorsi personalizzati per ciascun alunno; devono, tuttavia, essere analizzati a fondo i problemi etici legati all'uso dell'AI che fanno riferimento alla sorveglianza, alla privacy e alla possibilità di poter manipolare gli studenti. In riferimento a ciò, come analizza Neil Selwyn nel volume *Should robots replace teachers?*[1], è fondamentale che l'introduzione dell'AI non diventi un processo passivo e tecnocratico, ma una scelta didattica consapevole e sostenuta da riflessioni pedagogiche e critiche. In questo ambito i temi che afferiscono alla protezione dei diritti individuali e alla responsabilità sociale nell'uso delle tecnologie avanzate diventano focali. A livello nazionale troviamo pubblicazioni recenti quali *L'intelligenza artificiale a scuola. Guida per una pratica didattica consapevole*[2] di Giulia Lorenzoni e *La classe potenzIAta*[3] di Lorenzo Redaelli che propongono un approccio didattico in cui l'AI sia al servizio dell'insegnamento e non viceversa; questi contributi mostrano che l'AI non va introdotta come elemento neutrale ma va collocata in un contesto pedagogico e normativo che ne sostenga l'uso etico, responsabile ed equo.

2. IA e implicazioni etiche: esempi

L'IA presenta strumenti promettenti per personalizzare l'apprendimento e migliorare l'efficienza educativa, ma solleva anche preoccupazioni etiche, in particolare riguardo all'equità degli algoritmi che analizzano i dati degli studenti. Il rischio di discriminazione, causato da "bias" nei modelli algoritmici, è uno degli aspetti più critici di questa tecnologia. È fondamentale riflettere sulle implicazioni etiche dell'uso dell'IA nella didattica e specialmente su come gli algoritmi possano amplificare o creare nuove disuguaglianze tra gli studenti. I sistemi di IA utilizzati nelle scuole raccolgono una grande quantità di informazioni sugli alunni, comprese le loro conoscenze, lacune, progressi e persino abitudini di studio e dettagli biometrici, a seconda del programma di IA impiegato. Holmes[4]

[1] N. Selwyn, *Should robots replace teachers?*, Polity Press, Cambridge 2019.
[2] G. Lorenzoni, *L'intelligenza artificiale a scuola. Guida per una pratica didattica consapevole*, Lattes Editori, Torino 2024.
[3] L. Radaelli, *La classe potenzIAta.* Mondadori Università, Milano 2025.
[4] W. Holmes, M. Bialik, C. Fadel, *Artificial Intelligence in Education: Promises and implications for teaching and learning*, Center for Curriculum Redesign, Boston, 2019.

e Xu & Ouyang[5] evidenziano come la raccolta di dati, se non supportata da trasparenza e accountability, possa diventare uno strumento sotto il controllo di pochi attori e recare danno all'autonomia educativa. Lorenzoni, in Italia, suggerisce che la progettazione educativa includa attività di alfabetizzazione digitale preventiva, volta a far sì che docenti e studenti comprendano i protocolli di raccolta dei dati e i rischi ad essi associati. È necessaria una particolare attenzione e un elevato grado di cautela nella gestione e nel trattamento dei dati, prevedendo l'adozione di adeguate misure tecniche e organizzative volte a garantire la riservatezza, l'integrità e la disponibilità delle informazioni, nonché a minimizzare i rischi connessi a possibili accessi non autorizzati, alterazioni o perdite dei dati trattati. Gestire adeguatamente la privacy degli studenti è complicato poiché i loro dati sono spesso amministrati da terzi che controllano i database e le piattaforme di IA. Chi ha accesso a questi contenuti? Chi garantisce dai rischi di utilizzo improprio e/o non autorizzato delle informazioni? Le scuole devono bilanciare l'innovazione didattica con il diritto alla riservatezza degli studenti. La situazione è complicata dalla mancanza di normative uniformi tra i vari paesi e dalla scarsa trasparenza delle aziende tecnologiche riguardo all'uso dei dati; non è sempre chiaro quali dati vengano raccolti, il loro scopo e la durata della conservazione, rendendo difficile per le scuole effettuare verifiche accurate. Sarebbe opportuno proporre modelli di "data stewardship" partecipata, in cui le scuole e le famiglie possono controllare in modo trasparente i dati generati dagli studenti. Chi lavora nel settore deve essere consapevole che gli algoritmi, pur essendo progettati per personalizzare l'insegnamento, possono anche influenzare pensieri e comportamenti dei discenti. Gli esperti del settore si interrogano sulla neutralità delle piattaforme educative e su come l'IA possa influenzare le scelte degli studenti in modo inconsapevole. È essenziale affrontare questi interrogativi per prevenire manipolazioni, anche non intenzionali; le ricerche di Boulamwini e Floridi mostrano che il modo in cui gli studenti vedono l'intelligenza artificiale può influenzare quanto si aspettano da sé stessi e quanto si sentono motivati a imparare. Se la percepiscono come superiore o "perfetta", possono sentirsi scoraggiati o meno coinvolti.

Se gli algoritmi, a causa di pregiudizi, favoriscono determinate aree di studio o carriere, gli studenti potrebbero prendere decisioni future che non riflettono le loro vere aspirazioni, limitando le loro opportunità di crescita personale e professionale. Non esiste un algoritmo equo e neutrale; studi dimostrano che possono contenere pregiudizi, sia consapevoli che inconsapevoli, riflettendo i "bias" presenti nei dati di addestramento, il che può portare a trattamenti ingiusti, specialmente per studenti di minoranze etniche o gruppi svantaggiati. Se un algoritmo non considera adeguatamente le diversità culturali, socio-economiche e di genere, potrebbe ostacolare invece di supportare gli studenti nel superare le difficoltà. I sistemi educativi stanno iniziando ad utilizzare l'IA per raccomandare corsi, monitorare il rendimento degli studenti e/o valutarne le prestazioni. In una situazione come questa, il pregiudizio razziale può influenzare l'accesso alle opportunità educative e modificare la qualità degli strumenti proposti; ad esempio, gli studenti di colore potrebbero essere trattati in modo diverso rispetto ai loro coetanei bianchi a causa di pregiudizi storici che si riflettono nei dati, non giudicando esclusivamente le loro capacità o il loro impegno. Altro problema significativo è la discriminazione di genere: gli algoritmi possono rafforzare stereotipi, orientando gli studenti verso specifici percorsi accademici in base al sesso,

[5] W. Xu, F. Ouyang, *The application of AI technologies in STEM education: A systematic review from 2011 to 2021*, «International Journal of STEM Education», 9, 2022.

piuttosto che alle loro reali attitudini. Ad esempio, un sistema che consiglia corsi di matematica e scienze agli studenti maschi e corsi di arte alle studentesse potrebbe limitare le opportunità per le ragazze di sviluppare competenze scientifiche, mentre i ragazzi potrebbero essere esclusi da aree che richiedono sensibilità umanistica. Tale "bias", se non gestito adeguatamente, può perpetuare ruoli di genere tradizionali, limitando la libertà di scelta e l'equità nell'istruzione. Per questo motivo, è fondamentale che l'introduzione dell'IA nelle scuole non avvenga in modo isolato o puramente tecnico. Lorenzoni e Rivoltella, ad esempio, sostengono che l'IA vada inserita in percorsi didattici che non limitino l'apprendimento a un rapporto studente-macchina, ma che includano momenti di riflessione collettiva, valutazione critica, confronto tra pari e discussione etica e sociale, stimolando un apprendimento consapevole e partecipato.

Un ulteriore elemento critico riguarda il divario digitale, che rischia di amplificare le disuguaglianze nell'ambito dell'istruzione supportata dall'IA. Gli studenti provenienti da famiglie a basso reddito, infatti, potrebbero non disporre di dispositivi adeguati o di connessioni Internet sufficientemente stabili e veloci, limitando così la loro possibilità di accedere e utilizzare le risorse educative basate sull'intelligenza artificiale. Questa situazione potrebbe tradursi in forme di discriminazione indiretta, penalizzando tali studenti nella valutazione delle loro performance non a causa di competenze o risultati inferiori, ma per la carenza di reali opportunità di partecipazione alle attività online. Quando gli algoritmi analizzano i dati degli studenti, spesso non considerano le differenze individuali o le esperienze di vita ma solo i dati pregressi a cui hanno avuto accesso.

È dimostrato da numerosi studi che l'uso distorto dei dati e i "bias" nei sistemi di intelligenza artificiale possono avere effetti negativi in vari contesti, riproducendo e amplificando le disuguaglianze esistenti. Prendendo in considerazione, ad esempio, il modo in cui lavora l'algoritmo COMPAS, utilizzato nel sistema giuridico statunitense per prevedere il rischio di recidiva, si evidenzia come i pregiudizi tecnologici possano avere impatti concreti; COMPAS[6], infatti, ha mostrato risultati problematici, in particolare per la discriminazione razziale, penalizzando in modo sproporzionato le persone di colore. Un'analisi ha rivelato che il sistema classificava erroneamente gli imputati neri come ad alto rischio di recidiva, mentre sottovalutava il rischio per gli imputati bianchi. Questo esempio dimostra come i sistemi automatizzati possano rinforzare disuguaglianze preesistenti, riflettendo i "bias" storici dei dati utilizzati. Il caso di Amazon[7], invece, riguarda l'uso dell'IA nel reclutamento. Nel 2017-2018, l'azienda ha abbandonato un algoritmo progettato per selezionare candidati a causa di un "bias" di genere. Addestrato su curriculum di candidati passati, il sistema rifletteva la predominanza maschile nelle posizioni tecniche, favorendo i candidati maschi e scartando le donne, anche con qualifiche superiori. Questo episodio solleva interrogativi su come gli algoritmi possano riprodurre e amplificare disuguaglianze pregresse. In un contesto educativo, un algoritmo che valuta le performance degli studenti potrebbe facilmente commettere errori simili, discriminando le studentesse in settori dominati da uomini o viceversa. Sempre parlando di disuguaglianze di genere è

[6] J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks*, «ProPublica», 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Reuters, *Amazon ditched AI recruiting tool that favored men for technical jobs*, «The Guardian», 11 October 2018, https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

facile constatare la presenza di rappresentazioni stereotipate (e discriminatorie) con una semplice navigazione su Internet utilizzando motori di ricerca. Ad esempio, una ricerca di immagini associate alla parola "donna" mostra principalmente donne giovani e attraenti, mentre una ricerca di contenuti per la stessa parola chiave porta a risultati legati a cosmetici, moda o abbigliamento, se invece cerchiamo "uomo" il numero di contenuti relativi alla moda e le immagini di uomini giovani e aitanti è inferiore.

Studi sulla discriminazione derivante dall'uso dell'IA sono stati effettuati principalmente negli Stati Uniti poiché da più tempo utilizzano questi tipi di sistemi. Tra i vari esempi rilevanti, basato su studi americani, troviamo il sistema di concessione di prestiti adottato da diverse banche e finanziarie. Certi algoritmi di valutazione del credito penalizzavano minoranze etniche e residenti in aree svantaggiate[8]. Questi modelli, progettati per stimare la capacità di rimborso dei clienti, presentavano evidenti distorsioni razziali, poiché si basavano su dati storici influenzati da pratiche discriminatorie del settore bancario. Le informazioni utilizzate mostravano disuguaglianze nell'accesso ai finanziamenti tra gruppi etnici, con approvazioni inferiori per persone di colore, anche a parità di punteggio creditizio rispetto ai richiedenti bianchi. Tale esempio dimostra chiaramente come gli algoritmi predittivi possano aggravare disparità preesistenti e consolidare stereotipi sociali ed economici.

Anche in ambito sanitario possiamo dedurre che l'uso di algoritmi possa comportare il rischio di "bias", se, come è successo in Amazon, tali sistemi vengono addestrati su dati storici che riflettono squilibri preesistenti. Poiché nella medicina e nella sanità i ruoli di responsabilità sono stati storicamente occupati in maggioranza da uomini, un algoritmo di selezione o pianificazione potrebbe replicare e rafforzare questa disparità, penalizzando le donne e i gruppi sottorappresentati. Oltre al versante lavorativo, la letteratura mostra che "bias" emergono anche nell'erogazione delle cure: alcuni algoritmi clinici hanno infatti sottostimato i bisogni dei pazienti neri[9], riducendo l'accesso a programmi di assistenza. Ne deriva che l'introduzione non controllata dell'IA in sanità rischia di accentuare le disuguaglianze invece di colmarle.

3. Utilizzo dell'IA in contesti educativi

Nel contesto scolastico, diversi studi internazionali hanno evidenziato criticità legate agli algoritmi di correzione automatica, specialmente quelli basati su intelligenza artificiale per valutare test scritti. Un esempio riguarda gli esami di ammissione universitaria negli Stati Uniti, dove sistemi di valutazione automatizzata tendevano a penalizzare studenti con background linguistici diversi o provenienti da scuole non anglofone. Gli errori rilevati dagli algoritmi non derivavano da carenze nei contenuti, ma piuttosto da differenze stilistiche o linguistiche legate alla cultura di origine, gli studenti non madrelingua anglofoni vedevano ridursi le opportunità di accesso poiché non si conformavano agli standard linguistici con cui gli algoritmi erano stati addestrati.

[8] R. Browne, M. Sigalos, *A.I. has a discrimination problem. In banking, the consequences can be severe*, «CNBC», 23 June 2023, https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html.

[9] K. Manke, *Widely used health care prediction algorithm biased against black people*, «UC Berkeley News», 23 June 2023, https://news.berkeley.edu/2019/10/24/widely-used-health-care-prediction-algorithm-biased-against-black-people/.

Vanno segnalate, inoltre, criticità nell'uso scolastico di sistemi di IA per monitorare il comportamento degli alunni attraverso dati biometrici o interazioni digitali. Diverse analisi hanno messo in luce come questi strumenti possano sfavorire studenti con disturbi dell'attenzione o modalità di apprendimento atipiche. Ad esempio, software per il controllo dell'attenzione rischiano di classificare in modo scorretto studenti con ADHD come poco motivati o disinteressati, semplicemente perché utilizzano strategie diverse di concentrazione. Questo tipo di "bias" rischia di generare valutazioni distorte delle capacità e risposte educative non adeguate alle reali esigenze degli studenti.

Le scuole italiane che hanno iniziato ad utilizzare sistemi basati su algoritmi per personalizzare i percorsi formativi, monitorare i risultati e proporre contenuti su misura in base alle abitudini di studio degli alunni, hanno effettuato un'attenta progettazione senza la quale queste tecnologie rischiano di amplificare le disuguaglianze?

Alcune ricerche recenti[10] hanno messo in evidenza come i sistemi di apprendimento adattivo e le piattaforme scolastiche basate su intelligenza artificiale possano rafforzare le disuguaglianze già presenti. Se gli algoritmi non sono progettati con attenzione, rischiano di favorire studenti già avvantaggiati o con maggiori risorse, mentre chi parte da situazioni di svantaggio riceve un supporto meno mirato. In questo modo, l'uso dell'IA in ambito educativo, invece di ridurre i divari, può contribuire ad ampliarli, con effetti rilevanti sia nella scuola che nell'università. Algoritmi che non considerino la varietà di bisogni e contesti degli studenti, rischiano di escludere proprio coloro che richiedono più attenzione. Anche nelle università sono stati adottati strumenti predittivi di IA per individuare chi potrebbe abbandonare gli studi e avere quindi bisogno di un supporto; tali strumenti sono basati su dati storici come voti, presenze e attività extra. Queste analisi, spesso, riflettono pregiudizi legati all'origine sociale ed etnica, segnalando come "a rischio", con maggiore frequenza, gli studenti provenienti da famiglie meno abbienti o da minoranze, replicando disuguaglianze già radicate nel sistema educativo. Tra i principali rischi c'è il "bias" di genere, quando gli algoritmi riflettono stereotipi educativi già presenti nei dati storici, e quello razziale, come dimostrato dagli errori dei sistemi di riconoscimento facciale, meno precisi con i volti delle persone di colore. Se applicate anche a tecnologie educative, come il riconoscimento delle emozioni o il monitoraggio degli studenti, queste distorsioni potrebbero causare valutazioni scorrette, penalizzando alcuni gruppi e compromettendo l'equità dei processi educativi, influenzando negativamente l'accesso alle opportunità di apprendimento.

Un ulteriore esempio riguarda l'uso dell'intelligenza artificiale per valutare e finanziare le scuole. Alcuni algoritmi hanno favorito istituti in zone benestanti, penalizzando quelli situati in aree svantaggiate; basi di dati che ignorano fattori esterni, come il livello di risorse disponibili o il supporto sociale, hanno contribuito ad accentuare il divario tra scuole ricche e povere, riducendo ulteriormente il sostegno proprio dove sarebbe più necessario. Ad esempio lo studio *Algorithmic Bias in Education*[11] evidenzia come i sistemi educativi basati sull'IA possano perpetuare disuguaglianze, specialmente in

[10] W. Strielkowski, T. Veinbender, L. Volkova, O. Garanina, *AI-driven adaptive learning for sustainable educational development*, «Sustainable Development», vol. 33, n. 2, 2025. https://doi.org/10.1002/sd.3221; I. Molenaar, R. F. Kizilcec, B. Chen, *Unveiling the shadows: Beyond the hype of AI in education*, «Heliyon», vol. 10, n. 5, 2024, e30696.

[11] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, «OSF», preprint 2021, https://osf.io/preprints/edarxiv/pbmvz.

contesti rurali o con risorse limitate. Le scuole in aree svantaggiate potrebbero non avere infrastrutture adeguate per implementare efficacemente l'IA, portando a risultati meno accurati e a decisioni che non rispondono alle reali esigenze degli studenti.

4. Necessità del controllo degli algoritmi

I casi esposti dimostrano come i pregiudizi algoritmici incidano anche sull'istruzione, minacciando l'uguaglianza di opportunità. Per evitare che l'IA aggravi le disuguaglianze, è indispensabile progettare sistemi trasparenti, equi e inclusivi, vigilando attentamente sui dati utilizzati e correggendo eventuali distorsioni. Per questo l'introduzione dell'IA nelle scuole deve essere accompagnata da riflessioni etiche serie, evitando che strumenti pensati per supportare l'apprendimento diventino causa di esclusione. È essenziale che ogni algoritmo sia costantemente controllato e aggiornato, garantendo parità di trattamento per tutti, senza distinzione di origine sociale, etnica o di genere. È fondamentale che insegnanti e studenti sviluppino competenze critiche nei confronti delle tecnologie basate sull'intelligenza artificiale, per poter riconoscere e segnalare tempestivamente eventuali ingiustizie o distorsioni algoritmiche. Come osservano Williamson e Piattoeva[12], l'uso crescente di dati e algoritmi nelle pratiche scolastiche rischia di produrre decisioni opache, non sempre comprensibili da chi le subisce. In quest'ottica, la formazione digitale non deve limitarsi all'aspetto tecnico-operativo, ma includere anche la capacità di comprendere le logiche sottese ai sistemi di IA, per individuarne i limiti e le implicazioni etiche e sociali.

Servono regole chiare e rigorose, trasparenza sull'uso dei dati e un'informazione completa rivolta a famiglie e studenti sugli scopi e sui meccanismi degli algoritmi. Secondo Veale e Binns[13], i sistemi automatizzati, soprattutto se impiegati in ambiti delicati come l'educazione, devono essere progettati per garantire accountability e possibilità di contestazione. Questo significa che le decisioni algoritmiche devono essere comprensibili, verificabili e, se necessario, modificabili.

A ciò si aggiunge la necessità di garantire precise tutele sulla privacy e sul diritto alla gestione dei propri dati personali. Il *White Paper on Artificial Intelligence*[14] pubblicato dalla Commissione Europea sottolinea l'importanza di garantire agli utenti – soprattutto se minori – la possibilità di conoscere come vengono utilizzati i propri dati, oltre al diritto di rettificarli o cancellarli. In linea con questa posizione, Livingstone, Stoilova e Nandagiri [15] evidenziano come i diritti digitali dei bambini siano spesso ignorati anche nei contesti scolastici, e che sia fondamentale promuovere una cultura della responsabilità anche nella progettazione delle piattaforme educative.

---

[12] B. Williamson, N. Piattoeva, *Objectivity as standardization in data-scientific education policy, technology and governance. Learning,* «Media and Technology», vol. 43, n. 3, pp. 1-13, 2018, https://doi.org/10.1080/17439884.2018.1556215.

[13] M. Veale, M. Van Kleek, R. Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, «Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems ACM», 2018, https://dl.acm.org/doi/pdf/10.1145/3173574.3174014.

[14] European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, Brussels 2020, https://ec.europa.eu.

[15] M. Stoilova, S. Livingstone, R. Nandagiri, *Digital by default: Children's capacity to understand and manage online data and privacy*, «Media and Communication», vol. 8, n. 4, 2020, pp. 197–207, https://doi.org/10.17645/mac.v8i4.3407.

Solo combinando competenze critiche, regole trasparenti e garanzie solide sulla privacy sarà possibile costruire un ambiente educativo in cui l'IA sia uno strumento di supporto equo e responsabile, capace di valorizzare le potenzialità di tutti senza riprodurre o amplificare disuguaglianze preesistenti.

I sistemi di IA non raccolgono solo dati anagrafici, ma anche aspetti sensibili come comportamenti, preferenze di studio e dati biometrici. Questa mole di informazioni, spesso elaborata senza un controllo trasparente da parte degli utenti finali, pone seri interrogativi etici, soprattutto nel contesto scolastico dove gli studenti sono soggetti vulnerabili, soprattutto a causa della giovane età. I dati generati quotidianamente nelle interazioni educative possono essere utilizzati non solo per personalizzare l'apprendimento, ma anche per sorvegliare, valutare e classificare gli alunni in base a parametri non sempre chiari né scientificamente fondati[16] .

In particolare, le tecnologie di tracciamento dell'attenzione, il riconoscimento facciale e le piattaforme adattive rischiano di esporre gli studenti a una continua analisi comportamentale, potenzialmente lesiva della loro autonomia e del diritto alla riservatezza[17]. Inoltre, il trattamento automatizzato di tali dati può consolidare stereotipi e discriminazioni preesistenti se non viene costantemente monitorato da un punto di vista etico e sociale[18].

Solo un approccio condiviso e responsabile tra scuole, istituzioni e aziende può assicurare un equilibrio tra innovazione tecnologica e tutela dei diritti degli studenti. Come sottolinea il *White Paper on Artificial Intelligence* della Commissione Europea[19], le applicazioni di IA nel contesto scolastico devono rispettare principi fondamentali come la trasparenza, la non discriminazione, la protezione dei dati e la supervisione umana. È essenziale che i processi decisionali automatizzati siano accompagnati da meccanismi di controllo accessibili e da una governance partecipata, in cui le famiglie, i docenti e gli studenti stessi possano esercitare un ruolo attivo.

5. Normativa attuale e supervisione degli strumenti di IA

Per garantire un uso realmente etico dell'intelligenza artificiale nella scuola, è fondamentale considerare leggi attuali, innovazioni digitali e responsabilità educative. Il Regolamento Generale sulla Protezione dei Dati (GDPR) stabilisce che scuole e aziende devono raccogliere solo le informazioni indispensabili per scopi didattici, evitando dati superflui, anche se questi potrebbero migliorare la personalizzazione dei percorsi formativi. Le scuole, quindi, devono informare chiaramente studenti, famiglie e insegnanti sull'uso dell'IA, specificando finalità, dati coinvolti e funzionamento degli algoritmi. Allo stesso modo, le aziende produttrici di software educativi devono garantire la conformità alle

[16] B. Williamson, N. Piattoeva, *Objectivity as standardization in data-scientific education policy, technology and governance in Learning*, «Media and Technology», vol. 44, n. 1, 2019, pp. 64-76, https://doi.org/10.1080/17439884.2018.1556215.

[17] UNICEF, *Policy guidance on AI for children*, UNICEF Office of Research – Innocenti, 2021, https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf.

[18] M. Veale, M. Van Kleek, R. Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, «CHI 2018 Papers», https://dl.acm.org/doi/pdf/10.1145/3173574.3174014.

[19] European Commission, *White Paper on Artificial Intelligence*, cit.

norme e ai principi etici, così da consentire agli utenti di esercitare i propri diritti in modo consapevole. La protezione dei dati richiede soluzioni concrete, come crittografia avanzata, autenticazione sicura e aggiornamenti costanti, oltre alla formazione del personale scolastico nella gestione di informazioni sensibili. Solo così si possono ridurre errori e rischi, anche considerando le minacce informatiche. È essenziale che docenti e operatori scolastici comprendano l'importanza della sicurezza digitale e agiscano con attenzione.

Anche la fase di progettazione degli algoritmi è un tema cruciale: trascurare i valori etici potrebbe portare alla creazione di strumenti inadatti e dannosi per la scuola, capaci di generare discriminazioni o percorsi scolastici non adeguati. Gli sviluppatori devono quindi seguire criteri di responsabilità e imparzialità, ma resta aperta la questione su chi debba controllare e garantire tale imparzialità. Devono farlo le singole scuole quando decidono di adottare le piattaforme dell'IA, avvalendosi di esperti del settore, anche esterni, o l'amministrazione centrale validando, attraverso apposite commissioni di comprovati esperti del settore, i diversi strumenti di IA da utilizzare in ambito educativo lasciando agli istituti solo il compito di scegliere quale piattaforma "approvata" utilizzare? Se si volesse interpretare nel modo più ampio il concetto di "libertà di insegnamento" dovrebbero farlo le singole scuole, ma queste, anche consorziate, hanno a disposizione le risorse necessarie per far ricorso a professionisti veramente esperti del settore? Una risposta, per ora, non c'è, ma è chiaro che, in questo settore così delicato, non si può risolvere il problema affidando a personale interno all'istituzione scolastica, magari formato con un corso on line, il compito di effettuare una valutazione minuziosa di tutti gli aspetti degli algoritmi utilizzati e che impatteranno sulla qualità dell'educazione impartita alle future generazioni.

Diversi studiosi hanno messo in luce quanto sia importante affrontare con attenzione il rapporto tra intelligenza artificiale e scuola. La filosofa Vallor[20], ad esempio, sottolinea che non basta costruire algoritmi "etici per definizione": ciò che conta davvero è come questi strumenti vengono utilizzati nei contesti concreti, perché il loro impatto dipende anche dalle regole, dalle relazioni sociali e dagli obiettivi educativi. Vallor propone quindi un approccio che coinvolga non solo informatici, ma anche insegnanti, giuristi e sociologi, per garantire un uso responsabile della tecnologia fin dalla fase di progettazione.

Anche la sociologa Ruha Benjamin[21] invita a non affidarsi ciecamente alla tecnologia, soprattutto quando si tratta di decisioni che riguardano le persone, afferma che è necessario creare forme di controllo trasparenti e partecipate, simili a quelle usate nella medicina per valutare i trattamenti sperimentali. Se riportiamo queste riflessioni in ambito scolastico potremmo dire che le scuole dovrebbero coinvolgere anche studenti, famiglie e comunità locali nella scelta e nella valutazione degli strumenti di IA.

Anche altri studiosi si muovono su questa linea e propongono una vera e propria "governance" dell'intelligenza artificiale in ambito scolastico. Per loro non è realistico pensare che le scuole, da sole, possano controllare strumenti così complessi, né che tutto debba essere deciso centralmente dai ministeri. La soluzione, dicono, sta in un lavoro di squadra: con standard condivisi, verifiche indipendenti e più formazione per i docenti, così che possano capire davvero come funzionano gli algoritmi che usano.

[20] S. Vallor, *Technology and the virtues: A philosophical guide to a future worth wanting*, Oxford University Press, Oxford 2016.

[21] R. Benjamin, *Race after technology: Abolitionist tools for the new Jim code*, Polity Press, Cambridge 2019.

Infine, lo studioso Paul Prinsloo[22] mette al centro la responsabilità nella gestione dei dati scolastici. Ricorda che l'autonomia didattica è un valore importante, ma non può trasformarsi in solitudine decisionale. Quando si introducono tecnologie che incidono sul percorso formativo degli studenti, serve una vera cultura della cura, che tenga conto dei diritti, dei bisogni e delle differenze di tutti.

Da quanto detto finora, si evince che, in ogni caso, studenti e famiglie devono essere parte attiva del processo decisionale, conoscendo i rischi e le potenzialità dell'IA e contribuendo alla scelta degli strumenti utilizzati. Una corretta informazione e un dialogo costante tra scuole, istituzioni e aziende tecnologiche sono indispensabili per proteggere i diritti degli studenti e costruire un ambiente educativo sicuro e moderno. L'intelligenza artificiale applicata alla didattica si trova di fronte a un delicato equilibrio: da un lato, il principio del GDPR impone di limitare i dati raccolti, dall'altro, la personalizzazione didattica richiede grandi quantità di informazioni. Questo crea una sfida complessa, soprattutto per le scuole con poche risorse e competenze digitali limitate, che potrebbero persino rinunciare all'uso dell'IA per rispettare le normative. Oltre a regole precise, servono politiche pubbliche che supportino concretamente le scuole, tramite investimenti in tecnologia e formazione. Insieme al GDPR, normative come l'AI Act puntano a promuovere un'intelligenza artificiale affidabile, sicura ed etica, tutelando diritti fondamentali e interessi pubblici, tuttavia, le procedure per rispettare queste regole, rischiano di rallentare l'adozione dell'IA nelle scuole. Questa complessità, d'altra parte, può trasformarsi in un'occasione per cercare soluzioni nuove, capaci di garantire un utilizzo responsabile e sicuro dell'intelligenza artificiale nella didattica.

6. Possibili soluzioni

Le scuole potrebbero adottare in modo sistematico la valutazione d'impatto sulla protezione dei dati (DPIA), uno strumento prezioso per garantire che le informazioni personali siano gestite nel pieno rispetto delle normative vigenti. Questo processo tiene conto non solo dei rischi tecnici, ma anche delle implicazioni etiche, sociali e didattiche delle scelte compiute. Un approccio del genere favorisce uno sviluppo tecnologico che coniuga efficienza e responsabilità. Tuttavia, per adeguarsi davvero al GDPR, è necessario un cambiamento culturale all'interno delle scuole, che coinvolga in maniera attiva dirigenti, insegnanti, studenti e famiglie nella gestione consapevole dei dati. L'obiettivo non è solo rispettare le regole, ma anche costruire un clima di fiducia attorno all'uso degli strumenti digitali nella didattica. Le scuole, però, non operano isolate, ma fanno parte di un sistema più ampio, con forti implicazioni sociali ed economiche. È quindi fondamentale che partecipino attivamente alla definizione di politiche future, individuando anche eventuali disuguaglianze legate all'uso dell'IA e collaborando per ridurle. Se un domani fosse richiesto di valutare l'impatto sociale delle tecnologie usate in classe, la scuola avrebbe un ruolo centrale nell'evidenziare disparità e criticità.

In ambito educativo piattaforme affidabili, senza pregiudizi o discriminazioni, sono fondamentali per garantire standard minimi di sicurezza; le autorità di vigilanza dovrebbero garantire i controlli e monitorare il rispetto delle norme oltre ad offrire indicazioni pratiche alle scuole. In questo contesto complesso, la collaborazione tra scuole, enti di controllo e

[22] P. Prinsloo, S. Slade, *Ethics and Learning Analytics: Charting the (Un)Charted*, in C. Lang, G. Siemens, A. Wise, D. Gašević (a cura di), *Handbook of Learning Analytics*, Society for Learning Analytics and Research, 2017.

aziende tecnologiche diventa indispensabile, soprattutto per affrontare sfide delicate come l'uso di dati biometrici o sistemi di riconoscimento facciale per la gestione della sicurezza scolastica. Anche se queste tecnologie possono apportare numerosi vantaggi, vanno sempre valutate attentamente sotto il profilo etico. Inoltre, è essenziale che gli studenti, ma anche i genitori e i docenti, acquisiscano consapevolezza dei rischi legati all'IA e sviluppino le competenze necessarie per muoversi in una società digitale; tutti gli stakeholders devono essere protagonisti attivi del dibattito sull'uso della tecnologia e non semplici utilizzatori passivi, per questo è fondamentale integrare nei programmi scolastici l'educazione alla protezione dei dati, alla sicurezza informatica e all'uso critico delle tecnologie digitali. La tecnologia, infatti, deve restare uno strumento e non diventare il fine ultimo della didattica. Senza una riflessione attenta, si rischia di perdere il controllo e di generare nuovi problemi.

In definitiva, l'introduzione dell'IA nella scuola può offrire grandi opportunità, ma solo se accompagnata da solide basi etiche e giuridiche. Norme come il GDPR non devono essere viste come ostacoli, ma come garanzie fondamentali per proteggere i diritti di tutti e costruire un sistema educativo capace di far crescere persone preparate e consapevoli, pronte ad affrontare le sfide del futuro.

Mark Coeckelbergh, David J. Gunkel, *Communicative AI: a critical introduction to Large Language Models*, Polity Press, Cambridge 2024, 144 pp.

Giacomo Pezzano[*]

Questo libretto, tanto agile quanto denso, avveduto e documentato, si propone come una guida critica ai *Large Language Models* (LLM), capace di investigare sia il loro significato teorico, sia le loro implicazioni pratiche. Il suo punto di partenza è che la fine di Novembre del 2022 – data del rilascio di ChatGPT – segna non solo una trasformazione tecnica ed economica, bensì una vera e propria pietra miliare interna alla rivoluzione copernicana portata avanti dalle macchine dell'informazione. Per millenni, perlomeno nella tradizione che siamo soliti chiamare occidentale, gli esseri umani si sono definiti come gli unici esseri dotati di parola e linguaggio e su questa convinzione si sono erette le antropologie tradizionali – oltre che la filosofia stessa. Tuttavia, con l'avvento delle IA comunicative questo dato di partenza si trova a dover essere rivisitato, intaccando – se non distruggendo – il nostro senso di auto-identità e di eccezionalismo: abbiamo smesso di essere gli unici esseri discorsivi (p. 1).

A partire da questa constatazione, gli autori intraprendono un percorso suddiviso in sette capitoli, seguendo due direttrici fondamentali (pp. 2-3). Da un lato, le IA comunicative vengono viste come oggetti che richiedono il pensiero filosofico per discutere nozioni capitali come quelle di verità, comprensione, significato, coscienza, intelligenza, creatività, identità, autonomia, bias, privacy e autorialità. Dall'altro lato, le IA comunicative vengono considerate strumenti attraverso cui pensare, anzi ripensare, dando vita a una rivalutazione dei modi in cui la filosofia ha concettualizzato alcune tra le proprie questioni chiave. Si tratta di un libro certamente opportuno e che riesce sia a svolgere una funzione introduttiva, sia a offrire un contributo originale, risultando così utile tanto per chi studia quanto per chi fa ricerca. In questa recensione, ne presenterò estesamente i contenuti dividendoli in tre blocchi tematici: la prima sezione si soffermerà sul Capitolo 1, dedicato ad alcuni aspetti tecnici cruciali per comprendere il funzionamento dei LLM (§ 1); la seconda sezione ricostruirà i Capitoli 2, 3 e 6, focalizzati sull'insieme delle questioni etiche, giuridiche, politiche ed epistemiche legate all'utilizzo delle IA comunicative (§ 2); infine, la terza sezione si concentrerà sui Capitoli 4, 5 e 7, incentrati su un ripensamento radicale del nostro modo di intendere e praticare il linguaggio, la comunicazione e la scrittura (§ 3). *SPOILER ALERT*: il testo sarà chiuso da una breve nota sulla "scrittura assistita" alla base di questa stessa recensione (§ 4).

[*] Università degli Studi di Torino, e-mail: giacomo.pezzano@unito.it.

1. Dentro la scatola nera

Il Capitolo 1 ha come obiettivo di «demistificare le tecnologie dei LLM» (p. 11), seguendo una triplice articolazione: gli LLM sono situati nel contesto più ampio delle intelligenze artificiali che processano il linguaggio naturale (a partire dalle prime chatbot come ELIZA); ne vengono spiegate le operazioni tecniche e le applicazioni; se ne identificano alcune cruciali sfide tecniche.

Sui primi due fronti, gli autori insistono sull'importanza del passaggio dal paradigma simbolico a quello connessionista nelle ricerche sulle IA, associandolo a due differenti modi di trattare il linguaggio. Il primo rimanda alla pratica dell'analisi logica e grammaticale, in quanto approccia il linguaggio come un sistema logico-formale fatto di simboli e di regole. I segni verbali vengono etichettati sulla base della loro appartenenza a una categoria generale di cui rappresentano l'occorrenza (nome/verbo/articolo/preposizione), per poi venire combinati seguendo regole di assemblaggio predeterminate (articolo + nome + verbo + preposizione + articolo + nome). Il secondo – alla base delle correnti IA comunicative – segue una logica diversa, che approccia il linguaggio come un sistema di relazioni di occorrenza probabilistica, la cui mappatura è precisamente quanto viene demandato alle reti neurali artificiali del tipo *Generative Pre-trained Transformer*. Ma procediamo per gradi.

Sintatticamente, una sequenza del tipo "ieri sera ho mangiato un ___" risulta correttamente completata tanto dicendo "ieri sera ho mangiato un'astronave" quanto dicendo "ieri sera ho mangiato un asparago". Eppure, semanticamente, la prima suona decisamente bizzarra, mentre la seconda suona alquanto verosimile – indipendentemente dal suo effettivo valore di verità (effettivamente, non ho mangiato asparagi ieri sera). In breve, sulla base della sequenza data, la probabilità di co-occorrenza delle parole "mangiare" e "asparago" è più alta di quella delle parole "mangiare" e "astronave"; al contempo, le cose già cambierebbero se, per esempio, la sequenza iniziale fosse stata "ieri sera nei miei sogni ho mangiato un ___". Su queste basi, l'intero sistema linguistico può essere interpretato come una nebulosa dei vari coefficienti di probabilità che regolano le possibili combinazioni tra parole e sequenze di parole. Questo fa sì che il significato venga considerato in chiave endogena piuttosto che esogena: può essere calcolato su base puramente statistica, senza tener conto della relazione tra le parole e le cose o tra le parole e la mente.

È proprio qui che intervengono le reti neurali artificiali, incaricate di calcolare i vari coefficienti di probabilità per ogni possibile parola ed espressione nel modo più efficace ed economico possibile. L'operazione avviene nel seguente modo. Ciascuna parola riceve un "peso" sotto forma di valore numerico («vettore parola»), in modo tale da assegnare vettori simili a parole che condividono certe caratteristiche semantiche – così che, per esempio, "cane" risulta numericamente più vicino a "bassotto" e meno prossimo, ma comunque non troppo lontano, da "gatto". In questa maniera, ogni parola o sequenza ordinata di parole risulta circondata da un "alone" di termini o espressioni affini, una sorta di spazio di possibilità che – avendo mappato la rete di probabili dipendenze reciproche – consente di fare previsioni su quale termine o espressione preceda o segua un altro termine o espressione e così via. Tuttavia, c'è un ulteriore passaggio che interviene a rendere il processo più semplice ed effettivamente gestibile: anziché operare la predizione prestando attenzione a ogni singola parola, le reti neurali si avvalgono di un *attention network*, che

attribuisce alle parole-vettore anche un «valore di attenzione» tra 0 e 1, stabilito in base al peso che ciascun termine ha nella sequenza in esame. Questa ulteriore determinazione reciproca – la "trasformazione" indicata in *Generative Pre-trained Transformer* – genera un «vettore di contesto», che viene infine moltiplicato per il «vettore parola», così da aumentare il valore delle «parole-attenzione», cioè dei termini che rendono accurata la predizione, e diminuire quello delle altre parole (pp. 16-19).

L'intero processo di definizione dei coefficienti si basa su fasi iterative di addestramento (*pre-training*) e correzione retroattiva degli errori (*backpropagation*), che prevede miliardi di ripetizioni fino a quando il modello riesce a minimizzare l'errore tra la parola prevista e quella effettivamente presente nel testo di addestramento. Perciò, l'intera operazione risulta tanto più efficace quanto più – da un lato – vasto e vario è il dataset utilizzato e quanto più – dall'altro lato – si introducono livelli di trasformazione nella computazione, facendo sì che l'output di un livello diventi l'input di un altro e così via, secondo una scala di crescente complessità. Per esempio, se già GPT-3 possedeva 96 livelli, GPT-4 conta quasi 2 trilioni – *due miliardi di miliardi* – di parametri per attribuire i pesi distribuiti su 120 livelli di elaborazione: ecco perché sono modelli linguistici *large* (p. 20). Sulla base di questa architettura, le IA risultano capaci di generare testi non solo senza dover comprendere il senso delle espressioni verbali e senza sviluppare intenzioni comunicative, ma *proprio perché* prescindono da problemi di ordine semantico: una volta ricevuto un prompt, questo viene scomposto in *token* di 4-5 parole indipendentemente dal loro significato o contenuto informativo, prima di essere processato nei vari *layer* per arrivare a generare il responso.

È per questo che, venendo al terzo fronte, emergono problemi come i *bias* "assorbiti" dai dati di addestramento o le *allucinazioni* legate a stringhe testuali coerenti, ben formate e grammaticalmente corrette, ma che appaiono fattualmente sbagliate o fabbricate *ad hoc*, quando non prendono addirittura la forma di *nonsense* (pp. 22-23). Prima ancora di presentare implicazioni epistemico-cognitive (*overtrust* e autorità informativa), etico-giuridiche (copyright e responsabilità autoriale) o socio-politiche (manipolazione e governance democratica), queste sfide sono considerate come questioni prettamente tecniche, che chiamano in causa la necessità di soluzioni per perfezionare il *tuning*, introducendo, per esempio, criteri normativi e valori guida sotto forma di *feedback* aggiuntivi forniti da "controllori" umani o di principi che fanno da filtro automatico – stile la "Costituzione" adottata da Claude di Anthropic. Analogamente, le difficoltà poste dalla necessità di disporre di quantità massive di dati testuali e di ingenti risorse computazionali ed energiche per portare avanti l'addestramento vanno anche colte come questioni tecniche (pp. 24-25).

2. Quando l'apparenza fa la differenza

Le considerazioni più tecniche cominciano a cedere il passo a quelle più prettamente filosofiche con il Capitolo 2 (pp. 28-40), concentrato sui temi appena indicati e altri come la disoccupazione, il *digital divide*, il potere, l'*accountability* e la creatività. In questa prospettiva, si discutono i pericoli, a livello individuale come sociale, relativi alla diffusione di informazioni inaccurate anche se non deliberatamente menzognere, all'utilizzo massiccio di algoritmi di profilazione e decisione che riprendono e amplificano forme di marginalizzazione e discriminazione, alle crescenti asimmetrie e diseguaglianze tra chi – dai

lavoratori singoli a intere nazioni o gruppi di potere – trae vantaggio dalla presenza dei LLM e chi invece ne risulta svantaggiato o peggio. Emerge così la necessità di interrogarsi sulla responsabilità etica, sociale e politica, che richiede una maggiore attenzione alla distribuzione delle risorse e del potere, alla partecipazione democratica e al coordinamento internazionale non solo in fase di utilizzo, ma anche e prima di tutto in fase di design e sviluppo delle IA. In questo modo, può si possono affrontare a monte questioni come – tra le altre – lo sfruttamento ambientale, la protezione dei dati e la proprietà intellettuale, prima di avere a che fare con le loro conseguenze incontrollate a valle.

È su questo sfondo che i due autori cominciano a sollevare una questione che diventerà centrale nei capitoli 4, 5 e 7 (pp. 38-40). La proliferazione di testi generati da macchine che competono direttamente con quelli scritti da esseri umani genera criticità affatto peculiari: anche quando non si dà il caso di plagio o infrazione di copyright, si pone comunque un problema di autenticità e di trasparenza, ossia di paternità del testo. Queste pagine portano il mio nome, ma potrebbero non essere state scritte da me e non nel senso che potrebbero essere state invece scritte da un altro essere umano – il classico collaboratore o assistente del Professore: potrebbero essere state composte da una qualche IA generativa (vedi § 4). Come viene evidenziato nel libro, questa eventualità, a fianco degli aspetti etico-giuridici o etico-politici del rispondere delle opere che si mettono in circolazione, comporta anche un problema educativo: come fare in modo di salvaguardare l'esercizio della scrittura non-assistita, o perlomeno di progettare ambienti di apprendimento che favoriscano un rapporto critico con questa nuova tecnologia? Tutto ciò rimanda a un interrogativo ancora più radicale, che viene sollevato richiamando opportunamente il *Fedro* di Platone: che cosa significa scrivere? È un'abilità come un'altra, che può essere aumentata, o persino completamente appaltata, a una macchina? E come può cambiare il nostro modo di pensare – nonché di essere umani – quando il nostro modo di scrivere va incontro a trasformazioni tanto epocali?

Il Capitolo 3 (pp. 42-57) richiama il dibattito più tipico quando compaiono sulla scena le IA: sono esse realmente intelligenti e coscienti? Sono vere menti? Innanzitutto, gli autori osservano che le nostre definizioni di intelligenza, coscienza e mente sono piuttosto instabili e aperte, per poi richiamare l'esistenza dell'«effetto-IA», per il quale ogni volta che una capacità ritenuta "intelligente" viene replicata da una macchina, si tende a toglierle valore, rimuovendola dal dominio dell'intelligenza: questo evidenzia come le IA contribuiscano attivamente alla nostra definizione e *scoperta* di che cosa sia un comportamento propriamente intelligente. In secondo luogo, essi sostengono che la discussione sul carattere effettivo o fittizio dell'intelligenza macchinica è tuttora informata dalla distinzione – insieme ontologica, epistemologica e morale – tra realtà e apparenza risalente perlomeno all'allegoria della caverna platonica e che, per esempio, trova eco nel celebre esperimento mentale della stanza cinese di Searle. In alternativa a questo paradigma, la mossa teorica degli autori sta nel riscattare le apparenze dal loro statuto di inganno, per coglierne il valore interazionale e comunicativo, cioè la loro funzione costitutiva all'interno dello scambio sociale.

Tra le conseguenze di questo cambio di postura, troviamo l'idea per cui prendere sul serio questo carattere di intelligenza "apparente" richiede di riconsiderare il senso di eccezionalismo umano, per aprire alla possibilità che esistano altre menti e altri soggetti morali e giuridici. Da un lato, è vero che queste entità apparentemente intelligenti rimangono cose, artefatti disegnati e realizzati da esseri umani, così che attribuire loro una

personalità giuridica rischia di dare il lasciapassare al loro abuso deliberato in forma umana troppo umana. Eppure, dall'altro lato, esse non sono proprio come le altre cose, perché ci parlano e usano il linguaggio in forme che appaiono indistinguibili da quelle tipiche di una persona, intrattenendoci in interazioni sociali ed emotive. Pertanto, i LLM – osservano i due autori richiamando anche Esposito – ci sollecitano nientemeno che a mettere in discussione la dicotomia bimillenaria tra persona e cosa, su cui si fonda non solo ogni ontologia morale e legale, ma persino la nostra stessa auto-comprensione: per essere un'apparenza, è tutt'altro che irreale. In definitiva, il fatto di avere a che fare con macchine apparentemente linguistiche non deve portarci tanto a svelare il carattere illusorio di questa parvenza, bensì deve condurci a reinterrogare il nostro stesso modo di concepire il linguaggio.

Il Capitolo 6 (pp. 89-104) affronta il problema del rapporto tra apparenza e realtà da un'altra angolatura, stavolta legata all'accuratezza delle sequenze di parole generate dai LLM, per proporre una riformulazione in senso post-rappresentazionale del concetto di verità. Secondo questa, la verità non consiste in una corrispondenza oggettiva tra parole e cose, concezione fatta nuovamente risalire a Platone, o – perlomeno – il "gioco linguistico" a cui prendono parte i LLM non è quello constativo, basato sull'intenzione di rappresentare il mondo e descrivere stati di cose: rispetto a questo, essi non possono che figurare come una *bullshit machine*. Piuttosto, bisogna allargare lo sguardo, per considerare la verità come una *performance* sociale multi-stratificata, che dipende dalla relazione che si instaura tra partner comunicativi, sul piano dell'interpretazione come dei rapporti di forza. Ciò significa che ogni output dei LLM va sì valutato nelle sue conseguenze etiche e politiche, ma non semplicemente sulla base della sua correttezza fattuale, bensì per il suo effetto discorsivo, il quale non si limita al "rumore" che – per esempio – inquina una contesa elettorale, ma tocca la stessa formazione della mente del libero cittadino. Da ultimo, gli autori insistono sulla necessità di non cadere né in una fiducia cieca nei LLM, né in un loro rifiuto moralistico: ciò che serve è una responsabilità distribuita, in cui gli sviluppatori, gli utenti e i decisori pubblici condividano l'impegno per un uso consapevole, trasparente e situato, per stabilire a quale gioco linguistico LLM sono chiamati a (simulare di) prendere parte.

3. Il significato del significato

Con i capitoli 4, 5 e 7, veniamo al contributo più originale offerto dal libro – perlomeno dal punto di vista di chi sta scrivendo queste pagine. Infatti, essi raccolgono diversi spunti disseminati negli altri capitoli, per confrontarsi in maniera più serrata con il problema fondamentale posto dalle IA comunicative: che cosa significa propriamente parlare, scrivere e comunicare? Vale a dire: *che cosa significa significare?*

Il Capitolo 4 (pp. 58-73) distingue due modi di vedere il linguaggio e il significato e due modi di intendere la comunicazione. Cominciando con la prima distinzione, gli autori intendono mettere in discussione la visione del linguaggio come sistema segnico che si è andata consolidando sin da Aristotele: le parole scritte sarebbero segno delle parole parlate; queste, a propria volta, sarebbero segno del pensiero; questo, a chiudere la catena di rimandi, sarebbe segno delle cose. In breve, le parole si riferiscono alle idee che si riferiscono alle cose. Perciò, il linguaggio si presenta come uno strumento insieme comunicativo e rappresentativo per gli esseri umani – e soltanto per essi: serve per trasferire ciò che si ha in mente in un'altra mente, dunque per trasmettere e ricevere messaggi relativi

a determinati stati del mondo. Che le parole abbiano un significato non vuol dire niente di diverso da ciò – perlomeno secondo la concezione tradizionale, la quale – tuttavia – non esaurisce i modi di intendere il *logos*, come mostra in particolare la prospettiva (post)strutturalista e anti-logocentrica, da Saussure a Derrida, passando anche per il Wittgenstein dei giochi linguistici.

Infatti, secondo quest'ultima impostazione, il linguaggio non è un mezzo di cui i soggetti umani dispongono per enunciare stati di cose che trovano corrispondenza nella realtà. Piuttosto, esso va visto come un sistema differenziale in cui ogni significante prima e significato poi si configura sulla base della propria compagnia, vale a dire in relazione agli altri segni. Così come – poniamo – il suono "b" si determina relativamente al suono "p" e viceversa, il senso di "marito" si determina relativamente al senso di "moglie" e viceversa. Ne segue una semantica distributiva in cui i significati sono articolati contestualmente, di modo che il linguaggio stesso partecipa a stabilire la realtà mentale e mondana di cui si discorre, piuttosto che semplicemente comunicarla/rappresentarla. È questo che si intende con l'idea heideggeriana o lacaniana per cui "il linguaggio ci parla" in senso transitivo, ossia che siamo sempre anche giocati da esso, anziché soltanto giocare con esso: il *logos* dispone di noi, prima che essere noi a disporne. In definitiva, il linguaggio si comporta sempre come un co-autore dei nostri discorsi, non come il loro specchio trasparente.

Venendo alla seconda distinzione, la comunicazione può essere interpretata o secondo la visione trasmissiva o secondo la visione rituale. L'idea di trasmissione, anche in questo caso, risulta fortemente intuitiva, legandosi alla concezione del linguaggio come strumento espressivo in senso rappresentativo: comunicare vuol dire trasferire un contenuto mentale da un soggetto (il mittente) a un altro (il destinatario), passando attraverso un canale neutro rispetto al significato – input e output coincidono. Per la prospettiva rituale, invece, la comunicazione consiste in un'attività sociale reiterata e performativa, volta a creare e mantenere un senso comune all'interno di una comunità; perciò, essa non è semplicemente incentrata sulla condivisione di contenuti informativi, bensì è alle prese con operazioni di riconfigurazione e rimescolamento – di *remix*. Si pensi, per esempio, al *gossip* o al "cianciare" tra amici: ridondante, ripetitivo, disinteressato all'accuratezza, privo di linearità, e così via – al limite di apparire privo di un vero senso, al di fuori appunto della funzione relazionale e performativa di mantenere e alimentare il legame personale rimescolando e quasi frullando insieme i significati. La comunicazione qui si incentra proprio sulla messa in comune – sulla produzione del *communis*: input e output possono e, per certi versi, persino debbono de-coincidere.

Ora, secondo gli autori, queste distinzioni risultano cruciali per analizzare la linguisticità dei LLM: per sintetizzare, essi sono *macchine strutturaliste e ritualiste*. I LLM non esprimono qualcosa che sta nella loro testa e/o nel mondo, bensì mediano discorsi dentro cui si trovano situati, inserendosi in dinamiche di cui non hanno possesso. È dunque vero che i LLM sono "pappagalli stocastici" che non hanno la minima idea di che cosa sia un matrimonio, ma in questo "mimano" i comportamenti umani ben più di quanto si possa credere. Il fatto è che la produzione di significato è un'operazione costitutivamente distribuita, vale a dire che il senso è un effetto o un'emergenza – un prodotto frutto della relazione tra esseri umani e meccanismi "inumani", non un presupposto custodito nei recessi dello spirito. Al contempo, il gioco di libero "rilancio" e "reinterpretazione" tra input umani e output macchinici è in grado di generare uno scambio pragmaticamente significativo anche se i LLM non hanno nulla da comunicare in senso trasmissivo: persino

quando "allucinano", essi possono partecipare eccome alla comunicazione intesa in senso rituale, in cui il senso si genera anche in mancanza di intenzioni strettamente informative. Anche su questo fronte, gli autori invitano non tanto a chiedersi se i LLM pensino, quanto piuttosto a cercare un modo di pensare con essi – "insieme a" e "tramite" essi.

Questa impostazione anima anche il Capitolo 5 (pp. 74-87), incentrato sulla questione dell'autorialità e autorità, che riprende ed enfatizza la lettura strutturalista dei LLM – convocando questa volta anche Barthes e Foucault. Di fronte a un testo, viene spontaneo domandarsi chi ne sia l'autore, dunque anche quale sia la fonte della sua eventuale autorevolezza: con i testi generati artificialmente, il problema è che domande del genere appaiono svuotate di senso – perlomeno se approcciate con strumenti concettuali più classici. Infatti, ancora una volta, la sfida è revisionare la concezione che vede l'autore come il padre del testo in senso forte: come colui che è insieme origine, custode e garante del vero senso del testo, veicolo della sua intenzione di esprimere un determinato significato, che il lettore deve limitarsi a recepire, facendo il possibile per rimanere fedele alla voce primaria. In quest'ottica, a rispondere di un testo è sempre e soltanto chi lo ha generato, perché solo costui propriamente sa di che cosa esso parla. Evidentemente, i testi generati artificialmente si sottraggono a simile dinamica, risultando "senza autore" sia perché privi di un qualche pensiero alle spalle, sia perché spesso "non autorizzati", in quanto i materiali all'origine della rielaborazione vengono utilizzati senza il permesso da parte dei loro compositori umani. È una criticità giuridica, etica e politica, oltre che genuinamente letteraria: chi risulta responsabile di un testo "artificiale" che – poniamo – dà a qualcuno lo slancio decisivo per suicidarsi?

Anche in questo caso, gli autori prendono le distanze da risposte di tipo conservatore e difensivo – basate sulla semplice limitazione in ottica censoria. Piuttosto, si suggerisce un'attitudine più trasformativa, che fa perno sulla logica della co-creazione, ossia di un'autorialità ibrida e distribuita – una genuina co-autorialità, basata sul presupposto per cui la generazione di significato è un atto condiviso tra chi scrive e chi legge, chi produce e chi riceve, chi "autorizza" e chi interpreta, nel senso forte di attribuire significato *ex post*. Se questo è un tratto persino costitutivo della scrittura, ciò che fanno i LLM sarebbe semplicemente – si fa per dire – renderlo pienamente leggibile. Emblematica, in questo senso, è l'immagine del "remixer" o del "DJ testuale", che attinge da un archivio condiviso e ne riassembla i frammenti per generare nuove forme espressive, procedendo per campionamento e iterazione. Con ciò, si chiarisce a più riprese, non si tratta di romanticizzare l'IA, ma proprio di essere consapevoli che questa visione co-partecipativa richiede un'attenzione persino più mirata e acuta verso i rapporti di potere che strutturano l'uso e il controllo delle piattaforme basate su IA.

Con ciò, veniamo al Capitolo 7 (pp. 106-120), che conclude il percorso riprendendo la domanda di Flusser se la scrittura, intesa come mettere in fila segni alfabetici, possa ancora avere un futuro. Gli autori rispondono che siamo sì alla fine della scrittura, ma non della scrittura *in quanto tale*: piuttosto, stiamo assistendo al tramonto della sua concettualizzazione platonico-logocentrica. Essa fa della scrittura uno strumento subordinato al linguaggio, considerando la parola scritta come immagine sbiadita della parola orale – un surrogato della voce: arrivando quando il significato è già stato prodotto, il segno artificiale alfabetico risulta secondario rispetto al segno naturale vocale, nella misura in cui ne preserva e reitera sì la presenza, ma rendendola virtuale, dunque priva di anima e orfana. Sotto questo riguardo, i LLM sembrano radicalizzare simile scenario, presentandosi

come scrittura totalmente afona: non tanto perché non emettono suoni linguistici, quanto piuttosto perché danno forma a un'ulteriore tecnologizzazione della tecnologia alfabetica, comportandosi come segno artificiale digitale del segno artificiale analogico. Eppure, proprio su questo piano – si suggerisce – bisogna far valere la lezione decostruttiva.

Infatti, la decostruzione invita a denunciare la rimozione della funzione costitutiva che la scrittura – nonché le tecnologie cognitive più in generale – gioca nel dare forma al pensiero: ieri come oggi, il significato è frutto di una performance "umano-macchinica". Su queste basi, lungi dall'assistere alla fine della scrittura verbale, staremmo piuttosto assistendo a una fase di «iper-alfabetizzazione», una vera e propria «obesità di lettere», che – di contro alla denuncia dell'imminente pericolo per la nostra mente e persino umanità – va vista come un'opportunità di rivisitare il nostro modo di pensare alla scrittura e di scrivere sul pensiero, mettendo da parte l'ossessione per le intenzioni autoriali, la conquista del significato originario e la protezione della naturalità dei significati. Ciò impegna a rimarcare che il futuro della scrittura sta nella lettura, inteso come gesto di selezione e cura di ciò che bisogna leggere e di come bisogna leggerlo. Se non sappiamo chi parla, allora dobbiamo chiederci: per chi è questo testo? In quale contesto opera? Con quali effetti? È un problema non di origine, ma di destinazione: in questa formula può riassumersi, per assurdo, l'intenzione autoriale al centro di *Communicative AI*.

4. Chi ha scritto questa recensione?

Il libro comincia con una breve prefazione scritta da ChatGPT a modo di quarta di copertina e si conclude con le seguenti parole:

> È ragionevole che il lettore si chieda se il testo che ha appena letto sia il prodotto di un autore umano, un output generato su richiesta di un LLM o il risultato di una qualche forma di collaborazione uomo-macchina. Anche se offriamo i consueti token di autenticità – i nomi propri degli autori stampati sul frontespizio, una dichiarazione scritta che si tratta di "contenuto genuino al 100% generato dall'uomo", una filigrana o qualche altro certificato ufficiale – il fatto è che non c'è modo di saperlo con certezza.
> Questa indecidibilità è tipicamente considerata un problema da risolvere. Ma è proprio questa mancanza di certezza e questa ambiguità che rendono l'atto della lettura (e quindi il ruolo del lettore) potente, importante e interessante. La destinazione del significato e della comunicazione è importante almeno altrettanto, se non di più, della sua origine presunta o postulata retroattivamente. Siamo felici di ritirarci ora come autori; è il vostro turno. In risposta alle sfide lanciateci dalle recenti innovazioni nell'IA comunicativa, possiamo (e dovremmo) rispondere con fiducia con questa battuta finale: "Cosa importa chi sta parlando, ha detto qualcuno. Cosa importa chi sta parlando?".

Probabilmente, arriverà davvero il momento in cui, proprio come non ci chiediamo se per scrivere si è usato uno stilo, una penna d'oca, una BIC o un software di videoscrittura, non ci preoccuperemo di sapere se un testo è stato prodotto avvalendosi di una qualche IA. Per il momento, tuttavia, è forse più opportuno sciogliere ambiguità del genere – proprio al fine di contribuire a normalizzare la scrittura tramite IA. Perciò, dichiaro che nell'elaborare questa recensione mi sono avvalso di ChatGPT-4o: queste pagine sono dunque frutto di una scrittura distribuita ed estesa, in senso non solo agenziale, ma anche spazio-temporale (spalmata su diversi "qui e ora" diversi") e tecnologico (supportata da

differenti dispositivi). Si è dunque trattato di un processo di "scrittura assistita" sia nel senso che è stata aiutata da ChatGPT, sia nel senso che ha comportato una forma di supervisione da parte mia. In particolare, l'iter di lavorazione si è sviluppato come segue – nel suo scheletro:

- In una prima fase, già successiva alla lettura del testo in formato di *e-book*, ho impostato la chatbot con le indicazioni del tipo di lavoro che avremmo fatto e delle sue finalità.
- In una seconda fase, ho condiviso alcune note vocali di lettura sul libro in generale prima e su ciascun capitolo poi, domandando non di trascriverle tali e quali, ma già di rielaborarle in una forma più vicina all'espressività scritta che a quella orale.
- In una terza fase, nell'invitare a raccogliere, ordinare e provare a sistematizzare le varie parti del testo, ho condiviso alcune pagine selezionate del libro sotto forma di *screenshot* presi dall'*e-reader*, evidenziate in diversi colori per distinguere i differenti gradi di rilevanza dei vari passaggi. Ho poi chiesto di integrare il testo che stava prendendo forma alla luce di quanto si poteva vedere nelle pagine-immagine.
- In una quarta fase, ho ripercorso l'intera chat per estrarre i blocchi di testo più funzionali e rielaborarli in varie forme, mettendo insieme il testo finale della recensione. Nel fare ciò, mi sono direttamente confronto con il libro un'ultima volta, per rifinire il discorso.

Infine, ho condiviso il file di queste pagine, provando a richiedere una breve formulazione conclusiva, che fosse insieme efficace ma non eccessivamente retorica, in modalità *prompt injection* – vale a dire che le precedenti tre righe erano inizialmente scritte in bianco, risultando così leggibili alla chatbot ma non a un essere umano. Non senza qualche tribolazione e allucinazione nella ricezione di questo "prompt nascosto", la chiusura è dunque affidata alla parolaIA che mi ha accompagnato: *l'intelligenza artificiale non pensa, non parla, non comprende. Ma ci interroga. Ci costringe a rivedere cosa intendiamo per pensiero, per parola, per comprensione. E forse, in questo domandarci, qualcosa di nuovo comincia a pensare – dentro e oltre di noi.*

Nello Cristianini, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*, il Mulino, Bologna 2023, 216 pp.

Cristina Rebuffo[*]

«*La scorciatoia* di Nello Cristianini è un libro affascinante che esplora l'intelligenza artificiale e il suo impatto sulla società. Cristianini, un esperto nel campo, affronta temi complessi con chiarezza, spiegando come l'I.A. non solo automatizzi processi, ma prenda anche decisioni cruciali che influenzano la vita quotidiana. Il titolo del libro si riferisce alla "scorciatoia" presa dall'I.A. per risolvere problemi, bypassando la comprensione umana tradizionale. L'autore riflette criticamente su questa dinamica, interrogandosi sulle implicazioni etiche e sociali. Un saggio stimolante che invita alla riflessione sul futuro dell'umanità in un mondo sempre più dominato dalla tecnologia».

Il precedente capoverso è stato elaborato da ChatGPT, acronimo di *Chat Generative Pre-trained Transformer*, il più celebre, attualmente (Agosto 2024), *chatbot* basato su intelligenza artificiale e apprendimento automatico, specializzato nell'interazione testuale tra macchina e utente umano. Il testo è stato elaborato da questo *software* in risposta al semplice *input* «Scrivi una breve recensione su *La scorciatoia* di Nello Cristianini»: si è deciso di anteporlo alla recensione vera e propria in maniera forse un poco provocatoria, ritenendo che possa essere una curiosa guida per l'analisi del volume in oggetto e come esempio pratico dei suoi contenuti.

L'attività di ricerca di Nello Cristianini, professore di Intelligenza Artificiale presso l'Università di Bath, in Gran Bretagna, è volta, in generale, all'analisi su larga scala di contenuti multimediali (notizie e *social media*) e all'impatto sociale dei cosiddetti *Big Data* e dell'Intelligenza Artificiale stessa; nello specifico, l'obiettivo del suo saggio qui presentato è quello di porre importanti questioni in merito alla necessità di trovare delle vie, delle strategie, per imparare a convivere con le nuove tecnologie basate sull'I.A., limitando il più possibile rischi ed effetti collaterali derivanti dal loro utilizzo deregolamentato, improvvisato e non ragionato. L'autore ritiene, in particolare, fondamentale che, per rispondere a tali importanti questioni e per impostare una relazione proficua ma sicura tra individui, comunità umane e politiche e "macchine intelligenti", sia assolutamente fondamentale attivare un dialogo tra le scienze naturali, la scienza informatica e le scienze umane e sociali.

Non a caso, il volume si apre con il capitolo intitolato *Alla ricerca dell'intelligenza*, in cui viene esposta una interessante disamina del concetto stesso di "intelligenza", che noi attribuiamo tanto alle facoltà umane quanto a quelle di alcuni tipi di macchine e tecnologie

[*] Lessico di Etica Pubblica, e-mail: rebuffocristina@gmail.com.

di cui ChatGPT fornisce un esempio particolarmente noto e controverso. Qui, l'autore assume un atteggiamento critico nei confronti delle conseguenze derivanti dall'uso, appunto, del concetto di "intelligenza" da una prospettiva esageratamente antropocentrica, che ci porta ad assumere l'aggettivo "intelligente" come sinonimo di "somigliante a un essere umano"; al contrario, sottolinea l'autore, «dimostrare intelligenza non significa assomigliare agli esseri umani, ma essere capaci di comportarsi in modo efficace in situazioni nuove. Questa capacità non richiede un cervello […] non c'è un singolo modo di essere intelligenti […] È fuorviante attribuire qualità umane a tutti gli esseri intelligenti»[1], siano essi animali non umani o macchine. Un comportamento intelligente è, dunque, secondo questa definizione, un atteggiamento teleologico, in quanto consiste nella comprensione e nello sfruttamento dell'ordine dell'ambiente in cui l'individuo opera col fine di trovare una soluzione a un certo problema, e questo può essere fatto indipendentemente dalla presenza di un cervello umano, di una coscienza umana, di un corpo umano; semmai, a essere necessario per poter agire con intelligenza, è che l'ambiente in cui si agisce sia un ambiente "regolare", perché solo in presenza di regolarità è possibile, tanto per un individuo umano, quanto per un calamaro, quanto per un software, fare previsioni più o meno precise e certe sul futuro e dunque individuare la risposta più corretta a ogni situazione, ed è solo in presenza di regolarità che è inoltre possibile sviluppare esperienza e apprendimento. Ciò che propone Cristianini è, per sua stessa ammissione, una nuova "rivoluzione copernicana" che ci conduca finalmente a riconoscere intelligenze radicalmente alternative a quella umana, con l'obiettivo di «calibrare meglio quello che possiamo aspettarci da queste nostre creature»[2], che non saranno forse mai in grado di armonizzare una sinfonia per orchestra con lo stile di Beethoven o di comprendere se sia eticamente corretto utilizzare dati biometrici per scopi di lucro ma sono, ad esempio, in grado di interpretare un QRCode. Proprio a partire da questa "rivoluzione copernicana" circa la concezione dell'intelligenza si innesta l'idea della "scorciatoia" che dà il titolo al volume e che indica il *modus operandi* delle cosiddette intelligenze artificiali: ripercorrendo gli studi svolti negli ultimi decenni in merito al loro funzionamento, Cristianini perviene a indicare con tale termine il modo, alternativo o alieno all'intelligenza umana, con cui alcuni tipi di *software* sono in grado, a partire da regolarità statistiche presenti in un certo numero di dati o informazioni, di elaborare previsioni o soluzioni: questa scorciatoia è alla base, ad esempio, del funzionamento dei motori di ricerca sul Web o dei correttori automatici dei programmi di scrittura e messaggistica istantanea o dei suggerimenti personalizzati sui più comuni siti di *shopping online* o, ancora, della procedura con la quale ChatGPT ha elaborato il testo introduttivo a questa recensione. In tutti questi casi le intelligenze artificiali hanno però dovuto operare anche una seconda scorciatoia, relativa al reperimento di dati e informazioni in numero sufficiente per poter operare la prima, e una terza, consistente nell'apprendimento delle tendenze degli utenti a partire dall'osservazione dei loro comportamenti in risposta alle soluzioni via via proposte: ed ecco dunque che ritorna l'importanza, per l'agire intelligente, di un ambiente "regolare" a partire dal quale apprendere quale via seguire per proporre soluzioni sempre più precise ed efficienti. È questo il principio sul quale si fonda la tecnologia chiamata *machine learning*, che consente alle macchine di perfezionare le proprie abilità se non addirittura di acquisirne di nuove

[1] N. Cristianini, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*, il Mulino, Bologna 2023, p. 9.
[2] Ivi, p. 25.

senza aver ricevuto una specifica programmazione esterna, umana: «Il risultato sarà un comportamento originale, che potenzialmente il programmatore non avrebbe immaginato»[3], un comportamento che quindi potremmo definire, con l'Autore, "sovrumano", reso possibile dal fatto che la velocità e la mole delle esperienze maturate da una macchina sono del tutto estranee alle abilità umane.

Ora, alla luce di tutto ciò, è urgente chiederci «come possiamo assicurarci che le macchine non violino le fondamentali norme sociali, eseguendo "alla lettera" quello che viene chiesto loro»[4] o nel loro agire intelligente "sovrumano". Come possiamo, cioè, interagire con tecnologie che non sono solo teoreticamente ma anche eticamente aliene rispetto a noi? A partire dal capitolo 5 del volume, *Comportamenti imprevisti*, l'Autore inizia a evidenziare i potenziali (o reali) rischi ed effetti indesiderati provocati dall'uso indiscriminato delle intelligenze artificiali in molteplici ambiti nei quali se da un lato esse risultano estremamente efficienti nel pervenire a risultati attesi dagli utenti, dall'altro ignorano il rispetto di norme sociali ed etiche comunemente condivise. In merito a ciò, Cristianini fa presente che

> al momento, gli scienziati e i filosofi stanno esplorando diverse dimensioni della «fiducia» (*trust*): trasparenza, equità, responsabilità, accuratezza e verificabilità (o ispezionabilità). Con verificabilità (o ispezionabilità) intendiamo che ogni strumento software sia concepito sia dal principio in modo da poter essere facilmente ispezionato da terzi, per esempio un'agenzia istituzionale.
> Potrebbe essere una buona idea decidere che solo tecnologie «verificabili» o ispezionabili possano essere usate in settori regolamentati.[5]

La verificabilità dell'agire intelligente dei *software* di cui ci serviamo giornalmente, talvolta anche inconsapevolmente, sarebbe particolarmente utile, come sottolineato nel capitolo 6, per quanto riguarda *Messaggi personalizzati e persuasione di massa*, cioè in tutti i casi in cui le intelligenze artificiali vengono utilizzate non solo per analizzare e comprendere ma anche per influenzare i nostri comportamenti reali, compromettendo quello che l'Autore definisce il nostro "diritto all'autonomia". Ci si riferisce, qui, ad esempio, alle situazioni in cui i dati personali sia pubblici che privati degli utenti di un *social network* vengono utilizzati per inferire tratti del comportamento non solo degli utenti presi in esame ma del comportamento pubblico anche di altri utenti, con lo scopo di indirizzare loro in maniera mirata messaggi pubblicitari, consigli per gli acquisti, notizie più o meno vere, posizioni politiche e consigli elettorali, influenzando anche il comportamento reale di quegli stessi utenti. Sono molti i casi di cronaca, anche mediaticamente molto esposti, che hanno mostrato le conseguenze di questo tipo di utilizzo dell'I.A.; basti pensare alle vicende riguardanti Cambridge Analytica e al suo ruolo fondamentale per l'elezione di Donald Trump a Presidente degli Stati Uniti nel 2016, grazie a una pericolosa intersezione tra psicologia comportamentale, scienza dei dati e tecnologia *machine learning*, che ha mostrato che il cosiddetto "*targeting* psicologico" è in grado di influenzare il comportamento di masse di utenti tramite messaggi persuasivi mirati, costruiti sulla base di specifici bisogni psicologici della platea a cui si rivolge. In casi come questo è evidente che l'intelligenza artificiale non si limita a fornire un servizio ma manipola pericolosamente il comportamento virtuale e reale di masse di utenti, senza che questi ne siano consapevoli o

[3] Ivi, p. 75.
[4] Ivi, p. 83.
[5] Ivi, p. 97.

senza che abbiano fornito il proprio consenso, compromettendo pesantemente la loro autonomia nel prendere decisioni. Inoltre, l'utilizzo deregolamentato delle tecnologie descritte nella prima parte del volume, potrebbe generare effetti indesiderati anche in termini di benessere sociale e individuale derivanti dall'esposizione prolungata, il che può generare problemi di autocontrollo, polarizzazione affettiva, distorsione della realtà. Molto si è detto e molto si è scritto a tal proposito, negli ultimi anni, anche in ragione dell'uso prolungato delle tecnologie di ultima generazione durante il lungo periodo della pandemia da Covid19; tuttavia pochi sono ancora gli studi scientifici rigorosi che sappiano esprimersi in merito con chiarezza: di qui l'urgenza di studiare più a fondo la questione, date le sue possibili implicazioni.

Il volume si conclude così, con forza, con un'esortazione con cui si apre il capitolo 10, *Regolare, non spegnere*, che suggerisce che «non possiamo realisticamente ritornare a un mondo senza Intelligenza Artificiale, così dobbiamo trovare un modo di convivere in sicurezza con questa tecnologia»[6] e la ricetta per realizzare questo importante obiettivo si fonda, secondo Cristianini, su due principi-chiave: quello della responsabilità e quello della verificabilità:

> Decidere chi è responsabile per gli effetti di un sistema di IA sarà un passo cruciale: è l'operatore, il produttore o l'utente? E questo si lega al secondo fattore: la verificabilità, ovvero l'*ispezionabilità*. Come faremo a fidarci di sistemi che non possiamo ispezionare, a volte addirittura perché sono intrinsecamente costruiti in tale modo? Ogni ulteriore regolamentazione del settore dovrà stabilire fin dall'inizio che gli agenti intelligenti siano costruiti in modo tale da essere ispezionabili, e che quell'onere deve cadere sul produttore o sull'operatore. Avendo stabilito questo, sarà poi possibile discutere la loro sicurezza, equità, e tutti gli altri aspetti, che possono essere chiariti solo ispezionando l'agente.[7]

[6] Ivi, p. 187.
[7] Ivi, p. 196.

Silvia Dadà, *Vulnerabilità digitale. Etica, intelligenza artificiale e medicina*, Mimesis, Milano-Udine 2024, 256 pp.

Paolo Monti[*]

Il volume di Silvia Dadà, *Vulnerabilità digitale. Etica, Intelligenza Artificiale e medicina*, pubblicato da Mimesis nel 2024, si inserisce in un dibattito sempre più rilevante nell'etica contemporanea: quello sulla la riformulazione delle categorie morali fondamentali alla luce della transizione digitale che sta riguardando sempre più aspetti della vita umana.

Il libro prende le mosse dalla constatazione che la rivoluzione tecnologica non si limita a introdurre nuovi strumenti di azione e di cura, ma modifica in profondità le condizioni stesse della nostra vulnerabilità. L'obiettivo di Dadà è interrogarsi su un'etica capace di riconoscere la vulnerabilità digitale come tratto costitutivo della condizione umana nell'era dell'intelligenza artificiale e della nuova biomedicina. L'autrice si muove con competenza tra i campi della filosofia morale, della bioetica e della filosofia della tecnologia, con uno stile limpido e sistematico che riesce a tenere insieme il rigore concettuale e la sensibilità per i problemi concreti della cura, con particolare attenzione per i contesti di pratica clinica.

La tesi fondamentale del libro è che la vulnerabilità non debba essere interpretata come un deficit, una condizione di debolezza o dipendenza da superare, bensì come una condizione che definisce l'umano e che, in qualche misura, costituisce un presupposto stesso del pensiero e dell'azione morale. Seguendo un'intuizione che attraversa la filosofia della cura e la fenomenologia contemporanea, Dadà sostiene che l'etica nasce dal riconoscimento della nostra esposizione reciproca: la finitudine e la fragilità sono tratti che definiscono ogni relazione genuinamente umana e che sul piano etico costituiscono un appello alla responsabilità individuale e collettiva.

Tale consapevolezza si è da tempo fatta strada anche nel discorso pubblico e nella produzione normativa di alcune istituzioni internazionali. Per esempio, nella *Dichiarazione sulla bioetica e i diritti umani* dell'UNESCO (2005) la vulnerabilità è riconosciuta come elemento centrale della riflessione etica, sia in quanto condizione che in certa misura accomuna tutti i soggetti umani, sia come indicatore di speciali forme di fragilità ed esposizione al danno che caratterizza alcuni individui e categorie. Dadà riprende e approfondisce questa prospettiva, mostrando come la vulnerabilità sia al tempo stesso universale e situata: universale, perché inerente alla condizione umana; situata, perché assume forme differenti in base ai contesti sociali, economici e tecnologici.

Proprio su questo terreno va a collocarsi la nozione di vulnerabilità digitale, che rappresenta il cuore concettuale del volume. Essa descrive la forma assunta dalla vulnerabilità umana nel contesto di esperienze e forme dell'agire sempre più mediate da

[*] Università degli Studi di Milano-Bicocca, e-mail: paolo.monti@unimib.it.

dispositivi elettronici, logiche algoritmiche, reti di dati e sistemi di automazione. Non si tratta solo di minacce specifiche legate, per esempio, a un uso improprio delle informazioni personali o a un'invasione della privacy: la vulnerabilità digitale designa più ampiamente le forme che la vulnerabilità umana prende entro condizioni di strutturale di dipendenza da infrastrutture tecnologiche che mediano il nostro accesso al mondo, agli altri e a noi stessi. Una parte importante del lavoro di Dadà consiste dunque nel superare la visione strumentalista della tecnologia, ancora dominante nel discorso pubblico e, in qualche misura, anche nel dibattito etico. L'idea che la tecnica sia un mezzo neutro, il cui valore morale dipende esclusivamente dall'uso che ne facciamo, non è infatti più sufficiente a comprendere l'impatto profondo che le tecnologie digitali esercitano sui processi percettivi, cognitivi e relazionali.

In dialogo con la filosofia della tecnica di Don Ihde e con la filosofia dell'informazione di Luciano Floridi, Dadà sostiene che le tecnologie sono da comprendersi come trasformazioni della nostra esperienza corporea e ambienti di mediazione: esse riconfigurano la nostra esperienza del mondo e plasmano le forme della nostra interazione. La vulnerabilità digitale, dunque, non è solo un effetto collaterale della tecnologia, ma il modo in cui la nostra finitudine si declina e si amplifica entro nuove forme di mediazione. In questo senso, il testo propone una prospettiva fra l'etica della cura e la filosofia della tecnologia, rileggendo la responsabilità etica come capacità di rispondere all'altro in un contesto di interdipendenza complessa.

La parte conclusiva del volume si rivolge in modo specifico alle implicazioni delle tecnologie digitali e dell'intelligenza artificiale nei contesti di cura e medicina. L'autrice mostra come l'introduzione di sistemi algoritmici nella diagnosi, nella prognosi e nella valutazione dei rischi comporti un duplice movimento: da un lato, un ampliamento delle capacità umane; dall'altro, una nuova forma di opacizzazione e allontanamento della responsabilità umana nei processi di cura. I sistemi basati sull'intelligenza artificiale non partecipano dell'esperienza incarnata della vulnerabilità, dunque elaborano grandi quantità di dati, ma senza che questo comporti una "conoscenza" della fragilità che fonda il legame tra medico e paziente, tra professionista e assistito. Per questo la sua integrazione nella pratica clinica non può avvenire senza una riflessione etica che restituisca senso alla relazione di cura. L'IA è una tecnologia che cresce con velocità notevole e che viene declinata in strumenti specifici in contesti diversi: la domanda etica che sollecita non è dunque tanto quello generale di chiedersi se essa vada o meno impiegata, quanto quella di metterne in questione le garanzie di trasparenza, spiegabilità e partecipazione negli specifici ambiti di applicazione.

In ambito medico, la vulnerabilità digitale si manifesta sia nella dipendenza dei professionisti da sistemi informatici che condizionano le loro decisioni, sia nell'esposizione dei pazienti a procedure automatizzate che riducono la singolarità biografica a *pattern* statistici. In entrambi i casi, l'etica della cura deve fungere da stimolo riflessivo e da istanza di garanzia, riaffermando che la questione posta dalla dignità del paziente non è interamente riducibile a una analisi della sua condizione in base ai dati clinici disponibili e alle previsioni statistiche conseguenti, ma vada riconosciuta in una comprensione che è propria della relazione di cura.

Il contributo teorico più originale del volume è forse la formulazione del concetto di vulnerabilità digitale che Dadà propone come cardine di un'etica ripensata per l'era tecnologica. Assumere riflessivamente le implicazioni della vulnerabilità digitale richiede non solo di astenersi da un abuso delle possibilità tecnologiche che possano nuocere o interferire in modo improprio con l'autonomia e la dignità delle persone, ma anche di creare condizioni tecniche e materiali che permettano sia a chi agisce la cura sia a chi ne è destinatario di abitare la comune fragilità come esperienza di fioritura e realizzazione dell'umano. In questa prospettiva di condivisione consapevole, la vulnerabilità non giustifica paternalismi, ma fonda un'etica della cura abilitante: prendersi cura dell'altro significa metterlo in condizione di esercitare la propria libertà.

Da tale prospettiva derivano varie conseguenze etico-normative. Fra queste vengono presentate come particolarmente notevoli:

- La necessità per le istituzioni e i professionisti di perseguire una consapevolezza riflessiva della natura non puramente strumentale della tecnologia e di ciò che questo implica per le loro decisioni quotidiane;
- Una rilettura in chiave relazionale del concetto di autonomia all'interno dei documenti normativi che orientano la pratica nei diversi ambiti di cura;
- Una solidarietà tecnologica nelle relazioni di cura, fondata sull'accesso equo alle informazioni e su un coinvolgimento attivo, responsabile e consapevole dei destinatari della pratica clinica.

Dadà mostra così come il concetto di vulnerabilità digitale non indichi soltanto una caratterizzazione dell'esposizione al rischio sensibile alla dimensione tecnologica, ma sia più ampiamente una categoria etica produttiva, capace di fondare uno sguardo più consapevole nel solco dell'etica della cura.

Il libro si caratterizza per chiarezza nella trattazione di un dibattito ampio e articolato grazie a una scrittura priva di tecnicismi superflui e al tempo stesso rigorosa. L'autrice mostra una notevole padronanza delle fonti, dalla bioetica contemporanea (Beauchamp, Childress) alla filosofia della cura (Gilligan, Tronto, Kittay), dalla fenomenologia (Ricoeur) alla teoria dell'informazione (Floridi), riuscendo a farle dialogare in modo originale. In questo senso, un punto di forza del volume è la capacità di tenere insieme l'etica della cura e la filosofia della tecnologia, due tradizioni che troppo spesso non dialogano nel dibattito accademico. L'autrice, appoggiandosi su queste risorse filosofiche, articola una comprensione relazionale della vulnerabilità nei contesti di mediazione tecnologica che scongiura i tentativi di ridurre la responsabilità a mero adempimento procedurale.

La proposta che emerge in queste pagine cerca di evitare tanto il tecno-pessimismo quanto il tecno-entusiasmo, in favore di una via intermedia, che riconosce la vulnerabilità come condizione ineliminabile e la tecnologia come contesto in cui tale vulnerabilità si ridefinisce. A differenza di analisi bioetiche che si concentrano su casi-limite o dilemmi normativi, *Vulnerabilità digitale* costruisce un quadro sistematico che abbraccia un ampio spettro di pratiche di cura e contesti di assistenza ordinaria. In tutti questi ambiti, la persona non si presenta con i tratti di un'entità autonoma e isolata, ma come un «corpo vivente situato» la cui identità è strettamente legata alle condizioni biologiche, materiali e sociali che lo sostengono. In questa cornice di tipo relazionale, l'autrice prospetta un'etica normativa non riducibile solamente a regole deontologiche o a calcoli consequenzialisti,

scegliendo invece un approccio attento ai contesti specifici e orientato alla lettura del significato etico delle interdipendenze.

Se un limite può essere individuato, questo riguarda la dimensione socio-economica del potere tecnologico, che rimane un po' sullo sfondo. Le asimmetrie di capacità economica e di controllo, pur certamente rilevate nel testo, potrebbero essere più ampiamente tematizzate per mettere in evidenza il legame fra il piano esistenziale della vulnerabilità digitale con quello del suo legame con i rapporti di potere. Una maggiore attenzione al nesso tra vulnerabilità e disuguaglianza, ad esempio alla luce della critica femminista delle tecnologie o delle teorie postcoloniali del digitale, rafforzerebbe ulteriormente il quadro. In questo senso, una direttrice di ricerca che potrebbe essere sviluppata ulteriormente riguarda il rapporto tra vulnerabilità digitale e *agency* politica collettiva. Se la vulnerabilità è condizione condivisa, la consapevolezza delle sue implicazioni etico-politiche invita a mettere al centro i doveri collettivi che essa suscita all'interno della sfera pubblica.

Nel complesso, la forza del libro sta nell'offrire una visione integrata: non un'etica "applicata" alle tecnologie, ma una filosofia pratica dell'esistenza tecnologica, in cui la cura diventa il nome di una nuova responsabilità condivisa tra esseri umani e attori tecnologici. Dadà invita a riconoscere che non possiamo pensare l'etica come un insieme di regole esterne alle infrastrutture che abitiamo: siamo già, costitutivamente, dentro la rete delle nostre mediazioni. L'etica della vulnerabilità digitale, allora, oltrepassa il livello ingenuo della domanda sul "come usare bene" le tecnologie, ma fornisce piuttosto alcuni elementi di riflessione per alimentare il tentativo di "stare criticamente" in esse, comprendendone le logiche interne e a orientandole al servizio della dignità della persona umana.

Il testo, in conclusione, prospetta uno scenario che non è né distopico né utopico, ma profondamente realistico. Il corpo, nonostante i suoi confini, viene definito da relazioni che rendono la sua vita e la sua azione possibili. In questo quadro, la vulnerabilità digitale non è una minaccia da neutralizzare, ma si presenta piuttosto come un orizzonte profondamente umano entro cui costruire nuove forme di solidarietà e di cura. Per i filosofi morali, i bioeticisti e i professionisti della cura, *Vulnerabilità digitale* rappresenta un invito a ripensare la responsabilità nel tempo delle nuove intelligenze artificiali. Ma, più in generale, il volume ricorda a tutti che la sfida etica del nostro tempo non è proteggersi dalla tecnologia, bensì imparare a rileggere le categorie etiche e le norme pubbliche a partire dalla considerazione riflessiva che l'era tecnologica non annulla, ma per certi versi enfatizza, la nostra natura di esseri vulnerabili, interdipendenti, e per questo irriducibilmente umani, al di là delle singole forme prese storicamente dal progresso scientifico e tecnologico.