Empowering Vulnerability: Decolonizing AI Ethics for Inclusive Epistemological Innovation

Antonio Carnevale [*]

ABSTRACT: Recent studies reveal a convergence in the ethical guidelines of AI, emphasising the emergence of 'fundamental principles' for responsible AI. However, dissenting voices argue that these principles are insufficient to address the social impacts of AI, revealing a disconnect between ideals and implementation. This article indirectly explores the necessity of AI ethics. It delves into the complexity of cataloguing discriminatory biases generated throughout the lifecycle of AI systems, analysing various types of causal reasoning for discrimination: technical, counterfactual, and finally, constructivist/genealogical. From this exploration, the article derives two additional arguments. Firstly, a call to move beyond bias-based determinism as a singular approach to evaluating discrimination caused by AI systems, thereby recognising the influence of political and social dynamics, including strong appeals for AI decolonisation. Secondly, there is a need to reconsider advocacy actions for vulnerable subjects not merely as a mere claim of denied or marginalised identities but for their epistemic engagement with the world and with others. In this openness, where machine ethics also resides, vulnerability becomes a central epistemological construct to foster inclusive technological innovation, a decisive element in the context of the growing symbiosis between society and AI systems.

KEYWORDS: Discriminatory-sensitive bias; Algorithmic causality; bias-based determinism; AI justice; AI Decolonization; Vulnerability and empowerment.

SUMMARY: 1. Introduction: What AI ethics? – 2. Paper organisation – 3. Algorithmic bias and discrimination: a conceptual dilemma of causality – 4. Beyond a bias-based determinism: AI justice and decolonisation – 5. Empowering vulnerability – 6. Conclusions.

[*] *Researcher in Moral Philosophy, DIRIUM Department, University "Aldo Moro" of Bari; Co-founder of DEXAI – Artificial Ethics. Mail: antonio.carnevale@uniba.it. This work is partially supported by the project FAIR – Future AI Research (PE00000013), spoke 6 – Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU. The article was subject to a double-blind peer review process.*

1. Introduction: What AI ethics?

Studies comparing existing guidelines found that they converge towards the same principles, even more so in recent times[1]. This level of convergence suggests that we are arriving at a set of 'core principles', which is currently the most favoured approach towards principled RAI[2].
Although there is a niche trend to consider AI ethics as the correlate of a constructivist and socio-technical view of AI[3], approaches that posit that the *ex-ante* incorporation of moral principles – such as respect for human autonomy; prevention of harm; fairness; explicability[4] – into machine design is only one domain of articulation of ethics, broadly the philosophical-scientific debate is addressing a dual set of ideas. On the one hand, the necessity of tools to verify that the AI system actually respects the ethical values[5], and on the other hand, the related thought of framing the engineering of ethics in AI systems as an epistemological and practical issue rather than merely a matter of computer science causing[6]. For example, Morley et al. argue about the need to move from 'what' to 'how', that is, to close the gap between principles and practices by constructing a typology that may help practically-minded developers apply ethics at each stage of the machine learning development pipeline, and to signal to researchers where further work is needed.

But are we confident that this often-speculative rush to find methods and tools to verify how much ontologically and engineering-wise ethical principles incorporated into AI will epistemologically produce fairer, more equitable, and sustainable AI systems?

[1] A. JOBIN ET AL., *The Global Landscape of AI Ethics Guidelines*, in *Nature Machine Intelligence*, 1, 9, 2019, 389–399. https://doi.org/10.1038/s42256-019-0088-2 (last visited 29/11/2024).

[2] J. FJELD ET AL., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, Cambridge (MA), 2020. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420 (last visited 29/11/2024).

[3] M. ANANNY, K. CRAWFORD, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, in *New Media & Society*, 20, 3, 2018, 973–989. https://doi.org/10.1177/1461444816676645 (last visited 29/11/2024); A. CARNEVALE ET AL., *A Human-Centred Approach to Symbiotic AI: Questioning the Ethical and Conceptual Foundation,* in *Intelligenza Artificiale*, 18, 1, 2024, 9–20. DOI: 10.3233/IA-240034.

[4] These are the four ethical principles listed by HLEGAI (High-Level Expert Group on Artificial Intelligence), *Ethics Guidelines for Trustworthy AI*, Brussels, 2018-19. The principles rooted in fundamental rights, which must be respected to ensure that AI systems are developed, deployed and used in a trustworthy manner. «They are specified as ethical imperatives, such that AI practitioners should always strive to adhere to them. Without imposing a hierarchy, we list the principles here below in manner that mirrors the order of appearance of the fundamental rights upon which they are based in the EU Charter» (p. 11.). On an emergent consensus in the international milieu on these principles, see also: A. JOBIN ET AL., *op. cit.*; J. MORLEY ET AL., *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*, in *Science and Engineering Ethics*, 26, 4, 2020, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5 (last visited 29/11/2024).

[5] I. VAN DE POEL, *Embedding Values in Artificial Intelligence (AI) Systems*, in *Minds and Machines*, 30, 3, 2020, 385–409. https://doi.org/10.1007/s11023-020-09537-4 (last visited 29/11/2024).

[6] J. MORLEY ET AL., *op. cit.*; L. FLORIDI, *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford, 2023.

Indeed, some scholars radically argue that AI ethical principles are useless, failing to mitigate AI technologies' racial, social, and environmental damages in any meaningful sense. According to Munn[7], AI ethics are a gap between high-minded principles and technological practice. Even when this gap is acknowledged, and principles seek to be 'operationalised'[8], translating from complex social concepts to technical rulesets is non-trivial. In a zero-sum world, the dominant turn to AI principles is not just fruitless but a dangerous distraction, diverting immense financial and human resources away from potentially more effective activity.

The issue, however, is not just the abstractness and the high-minded principles of ethics but also its opposite: an excessive specialisation on certain aspects, as demonstrated, for example, in the debate on how to best incorporate the concept of 'transparency' in AI system design[9]. Similarly, others suggest that concentrating tightly on bias distracts us from more fundamental and urgent questions about power and AI. The moral properties of algorithms are not internal to the models themselves but rather a product of the social and political systems within which they are deployed. This means that AI ethics should be integrated with AI justice theories[10].

This ambivalence between overly abstract and overly specialised ethics can lead to a series of complications that may further complicate the already challenging governance of relationships between society, humans, and machines. One primary complication is the escalating tension and opposition between a 'hard' and 'soft' variant of digital ethics in AI systems.

As argued by Floridi, hard ethics typically involve discussions of values, rights, duties, and responsibilities – or more broadly, what is morally right or wrong, what should or should not be done – when formulating new regulations or critiquing existing ones. For instance, advocating for good legislation or aiming to improve existing legislation can be considered instances of hard ethics. Hard ethics played a role in dismantling apartheid legislation in South Africa. On the other hand, soft ethics operates within the same normative scope as hard ethics but considers what should or should not be done beyond existing regulations, not in opposition to them, or despite their scope, or to change them. In other words, soft ethics represents post-compliance ethics because the «obligation to do something implies the ability to do that something»[11].

While Floridi's analytical distinction ideally involves dialectical impulses, in reality, it is becoming increasingly marked by a stark and rigid opposition between abstraction and hyper-specialization. This anti-dialectical tension leads, on the one hand, to proposals of supererogatory ethics, meaning requests for something impossible, and on the other hand, to overly permissive ethics proposals, which

[7] L. MUNN, *The Uselessness of AI Ethics*, in *AI and Ethics*, 3, 3, 2023, 869–877. https://doi.org/10.1007/s43681-022-00209-w (last visited 29/11/2024).

[8] C. CANCA, *Operationalizing AI Ethics Principles*, in *Communications of the ACM*, 63, 12, 2020, 18–21. https://doi.org/10.1145/3430368 (last visited 29/11/2024); J. MORLEY ET AL., *op. cit.*; A. DYOUB ET AL., *Learning Domain Ethical Principles from Interactions with Users*, in *Digital Society*, 1, 28, 2022. https://doi.org/10.1007/s44206-022-00026-y (last visited 29/11/2024).

[9] M. ZALNIERIUTE, *"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism*, in *Critical Analysis of Law*, 8, 1, 2021, 139–153. https://doi.org/10.33137/cal.v8i1.36284 (last visited 29/11/2024).

[10] I. GABRIEL, *Toward a Theory of Justice for Artificial Intelligence*, in *Daedalus*, 151, 2, 2022, 218–231. https://doi.org/10.1162/daed_a_01911 (last visited 29/11/2024).

[11] L. FLORIDI, *op. cit.*, precisely see chapter 6.

serve to confirm or approve compliance with the existing law. This undermines the internal dynamism of ethics between politics and culture, that is, the faculty to move between strong adherence to or contestation of existing rules (politics) and their transformation based on self-regulation and social praxes (culture).

In such a condition of increasing indecisiveness and indeterminacy, ethics are not uncommon to be distorted and used maliciously. This represents a second type of complication. Indeed, the increasing presence of ethical guidelines, committees, and ethicists in both public and private sectors has led computer and data science researchers to question the role of 'ethics' in the tech industry. Critics argue that companies sometimes use ethics to deflect concerns about their behaviour or political crises. Additionally, ethics can be strategically employed to select principles that impose minimal limits on actions while appearing to contribute to the common good[12].

Finally, the degree of confusion and misleading applicability inherent in this state of division leads to a third complication, a growing frustration among stakeholders[13]. As Hagendorff notes[14], almost all the guidelines that have been produced to date suggest that technical solutions exist, but very few provide technical explanations. As a result, developers are becoming frustrated by how little help is offered by highly abstract principles when it comes to the 'day job'[15]. This is reflected in the fact that 79% of tech workers report that they would like practical resources to help them with ethical considerations[16].

Considering everything, do we really need an ethics of AI?

2. Paper organisation

Throughout this article, I will attempt to address this question indirectly. In the first part of my argumentation, I will examine how conceptually complex and challenging it is to catalogue discriminatory-sensitive[17] biases that might negatively cause alterations and harm in the design and development of

[12] L. Floridi, *op. cit.*; B. Green, *The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics*, in *Practice. Journal of Social Computing*, 2, 3, 2021, 209–225. https://doi.org/10.23919/JSC.2021.0018 (last visited 29/11/2024); G. Van Maanen, *AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics*, in *Digital Society*, 1, 2, 2022, 9. https://doi.org/10.1007/s44206-022-00013-3 (last visited 29/11/2024); B. Wagner, *Ethics as an Escape from Regulation. From "Ethics-Washing" to Ethics-Shopping?*, in E. Bayamlioglu et al. (eds.), *Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, Amsterdam, 2018, 84–89. https://doi.org/10.1515/9789048550180-016 (last visited 29/11/2024).

[13] J. Morley et al., *op. cit.*

[14] T. Hagendorff, *The Ethics of AI Ethics – An Evaluation of Guidelines*, in *Minds and Machines*, 30, 1, 2020, 99–120. https://doi.org/10.1007/s11023-020-09517-8 (last visited 29/11/2024).

[15] D. Peters, *Beyond Principles: A Process for Responsible Tech*, in *The Ethics of Digital Experience*, 2 May 2019. https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-ae-fc921f7317 (last visited 29/11/2024).

[16] J. Morley et al., *op. cit.*

[17] By 'discriminatory-sensitive', I refer to a range of specific quality requirements that, due to space limitations in this contribution, I would equate with (a) the seven ethical requirements defined by HLEGAI, *op. cit.*; and (b) the discrimination categories described in the volume by the European Union Agency for Fundamental Rights, European Court of Human Rights, & Council of Europe, *Handbook on European non-discrimination law*, Strasbourg, 2018. https://data.europa.eu/doi/10.2811/58933 (last visited 29/11/2024).

AI systems. Against this backdrop, I conduct an examination of various studies that have elucidated discriminatory causality according to three types of reasoning: technical, counterfactual, and constructivist/genealogical.

From this initial exploration of challenges, I derive two additional lines of argumentation to complement my contribution further. In the first of these lines, I argue that we must move beyond an exclusively bias-based determinism to evaluating AI systems' discriminatory-sensitive aspects. Behind each ethical dilemma and every algorithmic process leading causally to biased outcomes lies a complex web of political and social dynamics. These dynamics are influenced by pressing calls for justice, such as those advocating for AI decolonization, reshaping the understanding of causality to be fluid and relational rather than deterministic.

Secondly, within this intricate political landscape, the moral actions of humans and the operationalisation of trustworthy machines extend beyond the mere assertion or protection of marginalized and vulnerable identities or the adherence to binary oppositions. Rather, they represent an epistemic engagement with the world and with others, constituting a cognitive assemblage. AI ethics might play a pivotal role in illuminating and enriching this nuanced discourse, thereby shaping the landscape of digital innovation that lies ahead. Against this backdrop, my aim has been to discern a revitalized notion of vulnerability empowerment. This conception emerges as a central epistemological tenet driving inclusive innovation and ethical governance in anticipation and mitigation of the forthcoming symbiosis between society and AI systems.

3. Algorithmic Bias and Discrimination: A Conceptual Dilemma of Causality

One of the pivotal aspects for ensuring that AI ethics can genuinely transition from the theoretical-conceptual phase ('what') to the pragmatic-orientation phase ('how') is to find convergent epistemic approaches and evaluative measures concerning the thorny issue of *bias* and its *discriminatory causality*. It is now undeniable that AI, especially in variants involving the support of machine learning techniques or extensions of generative AI, inherently revolves around the theme of bias. In certain algorithmic programming paradigms, the practice of bias is not understood in a negative sense – as prejudice – but is used to indicate a 'deviation from a standard', which can, therefore, occur at any stage of the design, development, and implementation process[18].

If it is indeed true that a *design by-bias* cannot be entirely disregarded in the lifecycle of AI systems, how then can one distinguish biases applicable to design from those that may instead engender discrimination? Hence, identifying the causal reasons behind the discriminations produced by AI biases – even considering causality in a thick sense as something constructivist and genealogical[19] – is by no means trivial. The major problem lies in the polyvalent and multi-layered nature of bias manifestations, as they are identifiable (a) both in the replication and reinforcement of cognitive biases already present in historical world data and in those with a higher additional layer of direct responsibility

[18] L. Floridi, *op. cit.*

[19] See: I. Kohler-Hausmann, *Eddie Murphy and The Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination*, in *Northwestern University Law Review*, 113, 2018, 1163–1228; M. Ziosi et al., *A Genealogical Approach to Algorithmic Bias*, in *Social Science Research Network (SSRN) Electronic Journal*, 2024. https://doi.org/10.2139/ssrn.4734082 (last visited 29/11/2024).

stemming from interventions that may (b) unveil new associations, highlighting, with tangible results, connections and interdependencies among data never seen before, or (c) synthetically anticipate the formation of new biases, creating hypotheses of future realities that are currently unforeseeable.

In the following, I will consider various research and studies that have attempted to systematise discriminatory causal complexity through different theoretical and methodological proposals: discriminatory causality as (i) *technical reasoning, (ii) counterfactual reasoning, and* (iii) *constructivist and genealogical reasoning*.

Discriminatory causality as technical reasoning. At the groundwork of my inquiry, I posit discriminatory causality as the outcome of technical reasoning concerning the type of modelling and training of the AI system, particularly when employing machine or deep learning techniques. As Pasquinelli argues, within this type of causality, we must identify at least three levels: world, data, and algorithm biases[20]:

- *World bias*: in society, biases like race, gender, and class inequalities are already present, and datasets often reinforce these biases, perpetuating stereotypes. In this context, Crawford distinguishes between two types of harm caused by bias in algorithms: resource allocation harm, such as denying mortgages to minority groups, and social representation harm, like denigration or unfair classification based on race, gender, or class.
- *Data bias* occurs during training data collection, formatting, and labelling, often reflecting outdated and biased taxonomies that distort cultural and scientific realities. This bias becomes ingrained in machine learning algorithms, amplifying existing biases and distorting information further.
- *Algorithmic bias*, resulting from computational errors and information compression, exacerbates inequalities by distorting and amplifying biases present in both the world and the data. This distortion is akin to the anamorphic perspective used in art, where proportions are distorted to maintain shape. This illustrates how machine learning can magnify biases in unexpected ways.

While this approach is helpful in abstracting and defining analytical processes, it tends to overlook the social complexity of the real world[21]. This leads to a dominant mindset in algorithm development, characterised by 'algorithmic formalism', which is adherence to prescribed rules and forms[21]. One potential approach to mitigate this issue involves intentionally excluding certain specific data variables from the training of the algorithmic decision-making process. Indeed, the treatment of statistically relevant sensitive variables or 'protected variables', such as gender or race, is typically restricted or prohibited by anti-discrimination laws and data protection regulations, aiming to mitigate the risks of unfair discrimination. However, this type of intervention raises ethical questions at a higher level than technical reasoning, as we will explore in the subsequent types of discriminatory causal reasoning.

[20] M. Pasquinelli, *How a Machine Learns and Fails*, in *Spheres: Journal for Digital Cultures*, 5, 2019, 1–17. https://doi.org/10.25969/MEDIAREP/13490 (last visited 29/11/2024).

[21] B. Green, S. Viljoen, *Algorithmic Realism: Expanding the* Boundaries *of Algorithmic Thought*, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, 19–31. https://doi.org/10.1145/3351095.3372840 (last visited 29/11/2024).

Downloaded from www.biodiritto.org.
ISSN 2284-4503

Discriminatory causality as counterfactual reasoning. As some scholars have emphasised, this kind of reasoning has been found worthy of explanatory conjecture in Judea Pearl's theory of causality[22]. The author articulates causal complexity into three levels, titled 1) association, 2) intervention, and 3) counterfactual.

- The first level is called *association* because it invokes purely statistical relationships defined by the naked data. For instance, observing a customer who buys toothpaste makes it more likely that he/she buys floss; such association can be inferred directly from the observed data using conditional expectation. Questions at this layer are placed at the bottom level of the hierarchy because they require no causal information.
- The second level, *intervention*, ranks higher than association because it involves not just seeing what is but changing what we see. A typical question at this level would be: What happens if we double the price? Such questions cannot be answered from sales data alone because they involve changing customers' behaviour in reaction to the new pricing. Customer choices under the new price structure may differ substantially from those prevailing in the past.
- Finally, the top level is called *counterfactuals*, which is a typical question in "What if I were to act differently?" Thus, it necessitates retrospective reasoning.

Researchers have often applied this reasoning in AI ethics to understand whether a hypothetical intervention to alter a subject's protected characteristic would have changed the outcome[23]. Most notably, Galhotra et al. propose 'probabilistic contrastive counterfactuals', which help quantify a feature's direct and indirect effects on outcomes and provide actionable recourse to individuals negatively affected by such an outcome[24].

This type of reasoning benefits from providing an appreciable logical-argumentative framework within the field of explainable artificial intelligence (XAI), which aims to diminish the opacity of AI-based decision-making systems, enabling human scrutiny and trust. However, as argued by Kohler-Hausmann and Ziosi et al.[25], this model inclines to be flawed. In this way – Kohler-Hausmann claims – «discrimination is detected by measuring the 'treatment effect of race', where the treatment is conceptualized as manipulating the raced status of otherwise identical units (e.g., a person, a neighborhood, a school). […] The counterfactual causal model of discrimination is not wrong because we can't work around the practical limits of manipulation […]. It is wrong because to fit the rigor of the counterfactual model of a clearly defined treatment on otherwise identical units, we must reduce race to only the signs of the category, meaning we must think race is skin color, or phenotype, or other ways we identify group status. And that is a concept mistake if one subscribes to a constructivist, as opposed to a biological or genetic, conception of race. The counterfactual causal model of discrimination is based on a flawed theory of what the category of race references, how it produces effects in

[22] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge (MA), 2000. See also M. Ziosi et al., *op. cit.*

[23] Examples are provided by A.-H. Karimi et al., *Algorithmic Recourse: From Counterfactual Explanations to Interventions*, 2020, arXiv:2002.06278. https://doi.org/10.48550/ARXIV.2002.06278 (last visited 29/11/2024).

[24] See S. Galhotra et al., *Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals*, 2021, arXiv:2103.11972. https://doi.org/10.48550/ARXIV.2103.11972 (last visited 29/11/2024).

[25] I. Kohler-Hausmann, *op. cit.*; M. Ziosi et al., *op. cit.*

the world, and what is meant when we say it is wrong to make decisions of import because of race»[26].

Discriminatory causality as constructivist and genealogical reasoning. To avoid protected features like gender, race, disability, etc., being represented as discrete units, existing in isolation rather than in relation, computer scientists and AI ethicists should consider the frontline discussion in social sciences, in which, indeed, many studies are converging on the assumption that no one theory of causation satisfies all scientific domains or specific studies. Accordingly, one should construct an appropriate 'causal mosaic' for each research study to determine what is causally relevant and articulate one's assumptions and approaches for warranting one's causal claim(s)[27]. According to scholars like Ziosi et al., this explains why AI ethics needs to shift the focus to constructive and genealogical conditions rather than the consequences of discriminatory outcomes to emphasise the importance of understanding and preventing algorithmic discrimination. According to Kohler-Hausmann, «Discrimination is a thick ethical concept that at once describes and evaluates the actions to which it is applied, and therefore, we cannot detect actions as discriminatory by identifying a relation of counterfactual causality; we can do so only by reasoning about the action's distinctive wrongfulness by referencing what constitutes the very categories that are the objects of concern»[28].

4. Beyond a Bias-Based Determinism: AI Justice and Decolonisation

Technical and Counterfactual approaches are better suited for observing whether variables like gender, race, disability, etc., are independent factors rather than elucidating the specific role they play in comparison to other factors. However, *observing* a phenomenon does not necessarily equate to understanding it. Increasingly, AI ethics concerns itself with peering into algorithms with the aim of elucidating the opaque mechanisms surrounding inference operations and statistical distribution – to prevent well-known effects such as *over-* or *underfitting* – thereby enhancing the transparency of the AI system. Yet, transparency is a political construct and should not solely be sought *inside* the machinery, but rather, as Ananny and Crawford argue, *across them*: «The implicit assumption behind calls for transparency is that *seeing* a phenomenon creates opportunities and obligations to make it accountable and thus to change it. We suggest here that rather than privileging a type of accountability that needs to look inside systems, that we instead hold systems accountable by looking across them—seeing them as sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans»[29].

In other words, we require theoretical approaches and methodologies qualified for elucidating algorithmic causality beyond the intrinsic rationality inherent in their construction and programming. If we perceive the ethical quandary of a fair, equitable, and reliable AI to lie in rendering its 'statistical

[26] I. Kohler-Hausmann, *op. cit.*, here p. 1163.

[27] R.B. Johnson et al., *Causation in Mixed Methods Research: The Meeting of Philosophy, Science, and Practice*, in *Journal of Mixed Methods Research*, 13, 2, 2019, 143–162. https://doi.org/10.1177/1558689817719610 (last visited 29/11/2024).

[28] I. Kohler-Hausmann, *op. cit.*, here p. 1163.

[29] M. Ananny, K. Crawford, *op. cit.*, here p. 974.

unconscious'[30], so to speak, as transparent as possible, we risk confusion. Not only are we looking in the wrong place (within the algorithm rather than through it), but we also risk being ensnared by an «enchanted determinism»[31]. For a multitude of reasons, including the nonlinear trajectory from inputs to outputs, we have yet to develop a theory that can explain why deep learning techniques excel at pattern detection and prediction, leading us humans to assert claims about 'superhuman' accuracy and insight while remaining unable to fully explicate the origins of these outcomes.

In the essay *Empiricism and the Philosophy of Mind* (1956), Wilfrid Sellars, in his critique of logical empiricism, demonstrated that knowledge having foundations independent of the linguistic-conceptual dimension is a myth, namely the 'myth of the Given'. The logical empiricist reasoning goes roughly as follows: for knowledge to be meaningful and not merely a play of the intellect, it requires a clear grounding in the empirical realm. Knowledge must be founded on empirical grounds, which must be divorced from any intellectual operation or linguistic-conceptual act to fulfil their role as foundations. On the contrary, Sellars argued that empirical facts only play and can play the foundational role for knowledge because, from their inception, they exist within a specific linguistic and conceptual configuration.

Let us extend Sellars' thought to AI. The determinism we believe inherent in AI's ability to provide a plausible representation of a 'given' reality or even to predict its imminent historical occurrence is not ontologically significant in the strict sense, as it entirely lacks the foundational role played by empirical facts, instead offering regularities and evidence entirely stemming from a statistical configuration of knowledge. Thus, if its foundation lacks empirical facts from the bottom, its knowledge lacks a language that can be spoken, put into practice, externalized, understood, and misunderstood from above. It is a novel mythology, a determinism *doubly insignificant* from an ontological perspective.

Conversely, algorithmic determinism becomes significant when viewed through it, within the *political conditions of its sociotechnical possibilities*. This is what the most advanced studies in critiquing AI ethics, such as AI justice and AI decolonization[32], tell us.

On the one hand, AI justice help to reframe much of the discussion around AI ethics by drawing attention to the fact that the moral properties of algorithms are not internal to the models themselves but rather a product of the social systems within which they are deployed. A scholar like Zalnieriute argues, for example, that the current focus on AI procedural issues like transparency is blinkered, acting as an «obfuscation and redirection from more substantive and fundamental questions about the concentration of power, substantial policies, and actions of technology behemoths»[33]. According

[30] In this context, there are studies that have questioned how machines can have negative conscious experiences, as seen in: L. DUNG, *How to Deal with Risks of AI Suffering*, in *Inquiry*, 2023, 1–29. https://doi.org/10.1080/0020174X.2023.2238287 (last visited 29/11/2024).

[31] A. CAMPOLO, K. CRAWFORD, *Enchanted Determinism: Power Without Responsibility in Artificial Intelligence*, in *Engaging Science, Technology, and Society*, 6, 2020, 1–19. https://doi.org/10.17351/ests2020.277 (last visited 29/11/2024). See also K. CRAWFORD, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, 2021.

[32] See: L. MUNN, *op. cit*.; I. GABRIEL, *op. cit*.; S. MOHAMED ET AL., *Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence*, in *Philosophy & Technology*, 33, 4, 2020, 659–684. https://doi.org/10.1007/s13347-020-00405-8 (last visited 29/11/2024).

[33] M. ZALNIERIUTE, *op. cit*., p. 139.

to Munn[34], if ethical principles are situated within company cultures and broader systems of power, then it makes sense to expand the scope of ethical engagement. Or, put differently, if machine learning reflects, reproduces, and amplifies structural inequalities, then any ethical program must operate intersectionally, considering a wide array of social and political dynamics and questioning the «seductive diversion of 'solving' bias in artificial intelligence»[35].

On the other hand, decolonial theorists recognise parallels between territorial and structural coloniality in the digital era[36]. Digital spaces, akin to physical territories, are susceptible to exploitation and extraction[37], fostering digital-territorial colonialism. This extends to digital-structural colonialism, where colonial power dynamics persist through socio-cultural imaginaries and technological development rooted in unquestioned historical values. Data colonialism and capitalism theories acknowledge data as a resource exploited for economic gain, reflecting the coloniality of technological power. Algorithmic coloniality emerges as algorithms shape resource allocation, societal behaviour, and discriminatory systems, influencing labour markets and geopolitical dynamics[38]. Against this backdrop, Mohamed et al. propose introducing a decolonial foresight taxonomy[39]. It will identify sites of coloniality, such as algorithmic decision systems and ghost work, revealing structural inequalities with historical colonial roots. By recognising these sites, discussions on power and inequality in AI must acknowledge colonial continuities, ensuring a comprehensive understanding of the societal impacts of algorithmic systems.

5. Empowering Vulnerability

«We build material and electronic walls, fences, and dikes to keep out the viruses and dark waters of death. As technological beings, these are the sort of things we humans do. In fact, it is hard to imagine what our material culture would look like without the struggle against vulnerability: technology is our vulnerability guardian, and it is in the guardian's house that we live as technological, risk-phobic beings. We are vulnerable by nature, but we are also vulnerability-averse by nature. We are already rebels. We are the children of Prometheus»[40].

[34] L. MUNN, *op. cit*.

[35] J. POWLES, *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence*, in *OneZero (blog)*, December 7, 2018. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-inartficial-intelligence-890df5e5ef53 (last visited 29/11/2024).

[36] J. THATCHER ET AL., *Data Colonialism Through Accumulation by Dispossession: New Metaphors for Daily Data*, in *Environment and Planning D: Society and Space*, 34, 6, 2016, 990–1006. https://doi.org/10.1177/026377581663319 (last visited 29/11/2024).

[37] N. COULDRY, U.A. MEJIAS, *Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject*, in *Television & New Media*, 20, 4, 2019, 336–349. https://doi.org/10.1177/15274764187966 (last visited 29/11/2024).

[38] P. RICAURTE, *Data Epistemologies, the Coloniality of Power, and Resistance*, in *Television & New Media*, 20, 4, 2019, 350–365. https://doi.org/10.1177/1527476419831640 (last visited 29/11/2024).

[39] S. MOHAMED ET AL., *op. cit.*

[40] M. COECKELBERGH, *Human Being@Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, Dordrecht-New York, 2013, here p. 4.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

I shall begin by specifying that the thesis underlying this final part is not the mere, albeit non-trivial, observation that the empowerment of vulnerability signifies a marked interpretative and ideological shift from a perception of weakness, of fragility to be ashamed of, to one of human dignity to be protected (also with technological aids) and whose assertion makes us stronger—more comprehensively human. A series of studies, including disability, capability approach and feminist studies, have now placed the socio-political issue of vulnerability on this plane[41]. What I would like to highlight, however, is an *epistemic* nuance contained within the dynamics of vulnerability.

To be vulnerable is always to be 'vulnerable to something', something external. This means that being vulnerable describes a situation not inherently one of inferiority but of susceptibility to external inducements. However, let us investigate more closely what this 'being outside' of those things that make us vulnerable entails.

Let us begin by stating that the something to which we are vulnerable is not simply a brute natural fact external to us, which by its presence influences us in some way. That something is an event, it is something that not only lies outside but *comes from outside*. Consider seismic or environmental vulnerability, defined as the propensity to suffer damage because of inducements from an event of a certain intensity. Its mere presence is, therefore, not sufficient. What renders us vulnerable must also possess a certain intensity. Otherwise, the inducements would not trigger, and vulnerability would never transition – to borrow Aristotelian terms – from its nominal potentiality (vulnerability as a noun) to its practical actuality (being effectively vulnerable to that something, i.e., an attribute). Pushing further, one might venture an additional speculation. Precisely because being vulnerable is always 'being vulnerable to something', it could be argued that it is the intensity of external events – hence not the brute facts but the quality of events – that determines the type of inducement, which in turn determines the essence of vulnerability. This leads me to argue that vulnerability is not a causal condition but an epistemic openness to the world[42].

Nevertheless, 'coming from outside' is not the only possible direction of this openness. If we consider some emotional states of individuals, in addition to coming from outside, we must add a second and perhaps more important variant, which is *being put outside*. Indeed, those who are vulnerable are exposed, uncovered, sensitive, and easily hurt. A person with a vulnerable character is easily mortified, offended, or depressed. In this second variant, vulnerability is not an epistemic openness to the world, but to the relationships between oneself and others[43].

[41] Just to mention a few: S.G. Harding, *The Science Question in Feminism*, Ithaca, New York, 1986; D. Haraway, *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, in *The Transgender Studies Reader*, London, 2013, 103–118; Id., *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective*, in *Space, Gender, Knowledge: Feminist Readings*, London, 2016, 53–72; A. Carnevale, *Robots, Disability, and Good Human Life*, in *Disability Studies Quarterly*, 35, 1, 2015. https://ojs.library.osu.edu/index.php/dsq/article/view/4604 (last visited 29/11/2024); M.J. Haenssgen, P. Ariana, *The Place of Technology in the Capability Approach*, in *Oxford Development Studies*, 46, 1, 2018, 98–112. https://doi.org/10.1080/13600818.2017.1325456 (last visited 29/11/2024); D. Cirillo et al., *Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare*, in *Npj Digital Medicine*, 3, 81, 2020. https://doi.org/10.1038/s41746-020-0288-5 (last visited 29/11/2024).

[42] A. Carnevale, *Tecno-vulnerabili. Per un'etica della sostenibilità tecnologica*, Salerno-Naples, 2017.

[43] L. Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*, Durham, 2020.

It is precisely in this going back and forth from and towards the world and from and towards others that I see an epistemic aspect of vulnerability imbued. To echo the words of Coeckelbergh, cited earlier and which may now acquire a broader meaning, «We are vulnerable by nature, but we are also vulnerability-averse by nature. We are already rebels». Being vulnerable to something is neither an ontological condition (something unchosen that we find ourselves saddled with), nor normative (a habitus chosen by law) or rather, more accurately, vulnerability can be both things, but this will depend on the social and political choices we make and which will shape the levels of abstraction with which we define causal nexuses, including the discriminatory causality of algorithms. We are *not* rebels. We are *already* rebels. This implies that AI justice and AI decolonisation are not the politically strongest solutions to AI ethical weakness, but ways of posing the right social and political choices in order to produce different levels of abstraction, thus capable of governing the ethical issue of AI system opacity in a non-hegemonic and mono-ideological manner. And I mean *sociotechnical* opacity, which concerns not only machines but, as Hayles states, the 'cognitive assemblages' of our technosymbioses[44] or our techno-vulnerability[45]. «For example, deciding what areas of autonomy a self-driving car will have is simultaneously a decision about what areas of autonomy a human driver will (and will not) have. Such a system does not exist in isolation. It is also necessary to take into consideration the sources and kinds of information available for the entities in a cognitive assemblage and their capabilities of processing and interpreting it. Humans can see road signs in the visible spectrum, for example, but a self-driving car might respond as well to markers in the infrared region. It is crucially important to realise that the cognitive entities in a cognitive assemblage process information, perform interpretations, and create meanings in species-specific ways»[46].

6. Conclusions

So, revisiting the question posed in the introduction, which kind of ethics of AI do we need?

If we conceive of AI ethics as ensuring that a system, no longer produces biased outcomes – such as when a facial recognition program fails to identify the face of a person of colour – then we would argue against it. We do not require such ethics, as it fails to address the crux of the matter: since the system has the capacity for self-correction, what is needed are engineers who are more attentive and sensitive to revising datasets to include vulnerable individuals and social groups. Similarly, if we regard AI ethics as a rule-driven guideline toward hyper-compliance and meeting demands for greater transparency, explainability, etc., for instance, toward a corporation to disclose its algorithms, once more, we will say no, as well. Such ethics remains abstract, a *petitio principii*, as algorithms are in constant flux as the system learns, rendering transparency at one point means obscurity at another. Such ethics serve no purpose; it is far more advantageous to be supported by jurists and lawyers who at least have the framework of existing laws as a concrete perspective for regulation.

[44] N.K. HAYLES, *Technosymbiosis: Figuring (Out) Our Relations to AI*, in J. BROWNE, ET AL. (eds.), *Feminist AI*, Oxford, 2023 (1st ed.), 1–18. https://doi.org/10.1093/oso/9780192889898.003.0001 (last visited 29/11/2024).

[45] A. CARNEVALE, *Tecno-vulnerabili*. *Per un'etica della sostenibilità tecnologica*, cit.

[46] N.K. HAYLES, *op. cit.*, p. 14.

Conversely, if we conceive that AI ethics must have some minimal reference to *ethos*, the Greek term from which it originates, and which denoted ‘character’, signifying the guiding beliefs or ideals that characterize a community, nation, or ideology, then AI ethics must be relevant and attentive to at least two other aspects.

Firstly, behind every ethical challenge and every algorithmic process causing discriminatory biases, there exists a structure of political and social relations upon which strong demands for justice, such as those of AI decolonization, exert influence, rendering the framework of causality fluid, relational, and not the outcome of deterministic inference.

Secondly, any moral and advocacy actions of humans as well as any operationalisation of trustworthy machines happen within this socio-political openness and it is not merely a matter of claiming denied or marginalized identities, of a binary oppositional logic of black or white, but of epistemic positioning in the world and in relation to others, a cognitive assemblage that AI ethics can assist in bringing to light and colouring the digital innovation that is upon us.

In this openness, where machine ethics also resides, vulnerability becomes a central epistemological construct to foster inclusive technological innovation, a decisive element in the context of the growing symbiosis between society and AI systems.