Simona Tiribelli[*]

Inequalities and artificial intelligence

Abstract
This paper focuses on one of the most urgent risks raised by artificial intelligence (AI), that is, the risk of AI perpetuating or exacerbating unfair social inequalities. Specifically, this paper argues for the need of decolonizing ethical principles underpinning current AI design through relational theories in order to overcome the current limits of an oversimplified and mainstream mainly Western understanding of ethics in AI, which is hampering the design of AI systems as forces for a fairer and more just society.

Keywords
Artificial Intelligence; Inequalities; AI Ethics principles, Fairness; Decolonial AI

I. Inequalities in artificial intelligence

The issue of inequalities is particularly central in the debate in the ethics of artificial intelligence (AI). The centrality of the topic is due to a series of recent phenomena that have unveiled how AI systems can silently replicate and strengthen current and historical inequalities, by producing biased decisions in critical domains, from education and employment to criminal justice and healthcare. In this regard, it is worth mentioning the renowned case of the AI-based system used in the United States (US) criminal justice system for predicting recidivism, which has been shown to produce racially biased decisions for black people while favoring white indicted, as trained on historical data and therefore biased past sentences (Angwin et al. 2016); or the case of the AI system used in many US hospitals for determining access to special care programs, which has been shown to reify social disparities due to erroneous use of past medical expenditures, which are traditionally lower amongst black patients, as a rational proxy for determining patients-in-need priority order access (Obermeyer et al. 2019).

[*] Università di Macerata

Such controversial phenomena have spurred a large corpus of ethical literature focusing on what ethical principles and techniques should steer the design of such systems so as to prevent unfair outcomes and especially the perpetuation of systemic inequalities. On the theoretical side, more than 80 ethical frameworks of principles have been proposed to guide the design of such systems so as to make them more ethical, trustworthy, and fairer (Jobin et al. 2019). On the basis of such principles, a number of techniques have also been implemented mainly focusing on detecting and eliminating biases in the design and training dataset of AI systems. However, as it has been argued (Binns 2018; Selbst et al. 2019), such approaches have been shown to be insufficient in order to ensure fairness in AI, and specifically, to develop and deploy AI systems able to mitigate existing and new unfair social inequalities and thus actively contribute to achieve a fairer and more just society. Indeed, on the theoretical side, benchmarking AI ethics principles tend to be rarely adequately explored, especially via the lenses of moral philosophy, ending up being very often vague, as well as to reflecting mainly and/or exclusively a high-level and mainstream ethical theory.[1] As a consequence, on the practical side, such dearth of ethical depth often translates in a large number of technical tools aiming to foster fairness in and through AI by mainly guaranteeing AI systems to work in the *same* way for *all* the individuals[2], according to a "strict egalitarian" approach (Mittelstadt, Watcher, Russell 2023), which tend to level differences and foster only a formal equality, as that the law aims to ensure.

Such an approach to ethics principles in AI is particularly troubling, especially when it comes to prevent and/or mitigate unfair inequalities through the use of AI. Such an approach indeed provides only an *appearance* of ethics and risks to justify the *status quo* that is profoundly imbued with unfair social inequalities. What we are asking to such systems instead is not just to avoid replicating existing unjust inequalities with their use and outputs: we are asking them to truly help us mitigating them. In this aspect lies the main ethical justification for their large-scale societal use.

To contribute to such ethical goal, this paper aims to show the need of decolonizing AI ethics principles currently steering the design of AI in order to make it a force for more just and fairer societies. More spe-

[1] This lack of ethical depth and insights from moral philosophy leads AI ethics to be very often criticized to be toothless or useless (Rességuier and Rodriguez 2022), that is, high-level and ineffective, deprived to its substantial richness (its teeth) and reduced to a sort of "soft law", thus, easily becoming a tool used by companies for ethics washing (Bietti 2019).

[2] Consider, for example, fairness via "parity models", which aim to ensure equal AI performance and outputs for each member of the group considered.

cifically, we argue that current AI ethics principles are limited in order to mitigate and/or prevent morally problematic inequalities reproduced and/or exacerbated by AI, insofar as their understanding and conceptualization fail to account properly for the conditions in which the most vulnerable and historically marginalized live, who are also the most negatively affected by AI.

To this aim, in the following section, we criticize current ethics principles adopted in AI ethics as hindering the development of fairer AI systems to the extent they tend to be oversimplified and mainly reflect just a mainstream Western conception of ethics, that is geographically, culturally, and socially limited in scope for the design of AI used at a global level (Mhlambi and Tiribelli 2023). In particular, we argue how such oversimplification and limited understanding of ethics in AI leads to develop AI systems that neglect what shapes the social conditions in which individuals and the most historically marginalized make decisions and act, including *a priori* social injustices and systemic disparities affecting them. Finally, we provide a few insights on how to revise such principles via non-mainstream ethical theory such as relational ethical accounts, which are currently at the outskirt of the debate in AI ethics, and shed light on their value for the design of a fairer AI.

II. Inequalities and the limits of AI ethics

In the last decades, ethics has rapidly and extensively gained a central place in the debate on AI and algorithms. After a first wave of ethics mainly focused on speculating on long-term concerns raised by the use of AI systems, today the field is moving fast to develop conceptual and practical tools to make today's AI systems beyond more accurate increasingly fair and trustworthy. Such effort is visible by looking at the number of principled frameworks and ethical guidelines that have been proposed worldwide to develop AI systems so as to benefit society and especially promote fairness (Jobin et al. 2019). The majority of such ethical principles are mainly borrowed from bioethics (Floridi and Cowls 2019) and usually ask for the design of AI in a way that ensure the respect and promotion of *human autonomy*, especially in terms of deliberative decision-making and rational choice; *justice* and *fairness*, avoiding AI-based unfair treatment and outcomes, currently mainly operationalized via debiasing techniques, as well as *explicability*, that is, the intelligibility of increasingly complex and often opaque AI systems (Pasquale 2015), and machine-learning (ML) and deep-learning (DL) algorithms specifically, especially as they are used in socially critical and morally-loaded domains.

Although such principles play a critical role in the design of AI systems, as they set the ground of what should be respected as meaningful for us as individuals and our societies, in this section we argue that they are still very limited to steer the design of AI in a way that is truly ethically meaningful, that is, in order to truly benefit our societies that are deeply permeated by unfair social inequalities. We argue this thesis by expanding two main considerations. First, the widespread conceptualization of prominent AI ethics principles, beyond their high-level nature, tend to be highly oversimplified and this hampers our efforts to prevent AI to replicate or strengthen unfair social inequalities. Second, such simplified conceptualization mainly or exclusively reflects a mainstream Western understanding of ethics, which becomes problematic for the practical design, deployment, and use of AI in the real-life conditions of our diverse, plural, and multicultural societies.

II.1 AI ethics oversimplification

Let us expand our first consideration, that is, the *oversimplification* of AI ethics principles. To do so, consider, for example, one of the most widely acknowledged ethical principles in the field of AI: the respect and promotion of human autonomy. Such principle is particularly crucial in relation to the issue of inequalities, paradoxically, as currently formalized, even more than that of fairness. Indeed, the principle of fairness in AI ethics aims to ensure the development of AI in a way that does not discriminate people via its treatment and outputs in access to and to benefit from the opportunities it can generate (Floridi et al. 2018), and therefore, it considers the users more in their state of patients and/or beneficiaries, instead of pro-active agents. The AI ethics principle of autonomy, instead, asks to design AI systems in a way ensuring that they respect and boost people as final end-setters, that is, in their active capacity to express their decision-making power and agency (Floridi et al. 2018), and thus, to concretely enjoy, benefit from, and act on the opportunities raised by AI.[3]

However, as it has been pointed out (Prunkl 2022), the principle of autonomy underpinning prominent AI ethics literature and guidelines, though its crucial, tends to be very often oversimplified, resulting vague and sometimes opaque. Such oversimplification is clear if we consider the analysis carried out by Jobin et al. (2019) on globally benchmarking

[3] The import of promoting autonomy via AI should be clear: even if opportunities are distributed by AI fairly (according to the AI ethics principle of fairness as currently mainly formalized), some people might lack of the necessary power to express properly their agency so as to truly benefit from such opportunities (Tiribelli 2023).

AI ethics frameworks, where autonomy emerges as a key principle but is mainly understood in limited and highly vague terms, such as "self-determination", "informational self-determination", and/or "privacy-preserving human control and oversight" on AI; or as "freedom to withdraw consent" or "freedom from exploitation, manipulation, and surveillance" (Jobin et al. 2019, p. 11; see also HLEGAI 2019). Beyond its many definitions, it sounds questionable that the respect and promotion of human autonomy can be reduced to maintain people's full control over themselves and AI (Floridi et al. 2018; HLEG-AI 2019; Fjeld et al. 2020), as well as to the exercise of informed consent via privacy techniques (European Parliament 2017; IEEE 2017; WHO 2021). Such ethical oversimplification is due to the fact that the philosophical complexity and the cultural richness of a key ethical concept as that, in this case, of autonomy emerge as poorly explored in AI ethics. Such a dearth of in-depth ethical inquiries on AI ethics principles is problematic. Indeed, in this case, the lack of a multilayered and multidimensional conceptualization of autonomy hampers a proper ethical understanding of all the diverse risks that AI can raise to human autonomy, and especially the autonomy of *whom* is mostly impacted by such systems, thus obscuring potential differences in people's valuing, experiencing, and exercising autonomy. As a consequence, this oversimplification hinders the design of AI systems that can effectively promote individuals' autonomy in truly adequate ways, which also means in different social conditions across diverse geographies, as well as according to different meanings that autonomy assumes in heterogeneous cultural contexts, given the transnational nature and application of AI.[4] This leads us to our second consideration. Indeed, beyond the necessity of a proper ethical understanding of such principles in order to clarify what they truly demand in our globalized societies, another consideration we should raise is the autonomy of *whom* such definitions of autonomy aim to and can effectively preserve.

II.2 Mainstream Western ethics for global AI

The second critical consideration we focus on in this paper concerns the representativeness of such AI ethics principles of the people subject to AI systems, that is, the people such principles should protect and empower through the use of AI. Following our previous example on autonomy, an accurate analysis of the main frameworks and initiatives (Jobin et al. 2019) and literature (Mittelstadt et al. 2016; Floridi et al. 2018; Milano et al. 2020; Calvo et al. 2020; Floridi and Cowls

[4] Similar concerns have been expressed for other AI ethics principles, such as for the AI ethics principle of fairness (see Giovanola and Tiribelli 2022).

2019; Tsamados et al. 2022) shows that autonomy is mainly or exclusively understood in AI ethics via the lens of traditional or mainstream Western philosophy, according to which, even if with some variations, autonomy is mainly grounded on the individual's capacity for rational deliberation and choice among alternative options, which in turn expresses individuals' capacity of self-governance and control via the exercise of reflective endorsement on their own reasons, preferences, and beliefs. Such a concept is limited insofar as it is mainly confined to a Western liberal (Kantian-inspired) understanding of autonomy, rooted in rational deliberation and decision-making, ignoring other precious contributions offered, for example, by feminists and relational scholars broadly. Indeed, even if such a concept has been widely encompassed in Western moral philosophy, it has been strongly criticized too. Such liberal understanding of autonomy and its focus on competencies such as rationality has been widely criticized in moral philosophy as unable to account for the social and cultural features informing and shaping people's agency conditions and their identity, and therefore, to properly account for people living in oppressive social conditions of various kind, such as those affected by epistemic injustice (Fricker 2007) and/or socio-economic and health constraints (Jaworska 2009).

In this regard, scholars in AI ethics have started to criticize such understanding of people as rational decision-makers which underpin AI design (Dignum 2022) and AI ethics – on which also this emerging mainstream understanding of autonomy rests. Their main argument is that designing AI systems according to such a rational model is at odds with the way in which the majority truly chooses in real-life conditions of "bounded rationality" (Thaler and Sunstein 2009; Kahneman 2011; Simon 1991). Moreover, this rationality-driven approach is problematic, as it veils behind illusory objectivity those unfair asymmetrical and hierarchical power dynamics of colonial heritage that silently nurture and subtly perpetuate systemic injustice and structural inequalities in our societies (Birhane 2021). For this reason, such scholars invite us to shift from a *rational* to a *relational* approach to AI (Dignum 2022; Birhane 2021) and we claim here to AI ethics too. Such a shift asks to move away from a paradigm of rationality, which is often efficiency-driven (objective cost-benefit analysis) and usually implies resource exploitation, from users' personal data to people themselves (Zuboff 2019; Couldry and Mejias 2019), to embrace relationality. By stressing the importance of focusing on relational and social aspects shaping our identity and agency conditions, relationality allows us to center the design of AI on the experiences of people that are mostly marginalized and vulnerable, and to date, the most negatively impacted by AI (Birhane 2021; Mohamed et al. 2020). Indeed, power asymmetries and structural inequalities happen and are embedded in so-

cial and relational practices and contexts. As a consequence, a relational shift to AI means designing AI systems that can help to discover and compensate for morally wrong and unfair historical power asymmetries and inequalities by investigating the relational contexts and social practices in which they arose, develop, and perpetuate (Mhlambi and Tiribelli 2023; Tiribelli 2023).

Centering AI ethics design on the most marginalized and the most vulnerable, such as people historically racialized and affected by systemic and epistemic injustices, amounts to decolonizing AI ethics. As Mohamed et al. (2020, p. 664) pointed out, adopting a decolonial approach in AI means putting the most marginalized and vulnerable people "who continue to bear the brunt of negative impacts of innovation and scientific progress" at the center of the design of such technology. Decolonizing AI means discovering how AI systems can replicate and exacerbate those systemic and structural harms and oppressive logic produced by colonialism and mitigate them. Decolonizing AI ethics means in turn acknowledging what AI ethics principles might obscure such asymmetrical power logic and revising them to make AI a force to dismantle historically rooted inequalities. As it has been shown (Jobin et al. 2019), AI ethics discourse is today mainly or exclusively shaped by US-European countries, while non-Western voices and geographical areas such as Africa and South and Central America are deeply under-represented. This scenario explains the mainly Western approach to AI ethics principles. However, this is problematic as it reinforces the disparities between those who have or do not have voice and power of agency to shape AI technology and according what idea of good and of a good society – ideas that can vary across cultures.

To sum up: it sounds that if we aim to not only avoid AI reinforcing unfair historically rooted power asymmetries but also to use AI to mitigate them and promote fairer societies, we are called to revise current AI ethics paradigms with alternative views, especially those of the most marginalized, as asked for by relational and decolonial approaches, namely, with the views of those situated mainly at the outskirt of Western Euro-American tradition and in non-Western perspectives, by centering such views and voices in AI research and design practices.

III. Decolonizing AI ethics via relationality

In the previous sections, we have argued that if we want to prevent AI from perpetuating or exacerbating current and historical inequalities, we have to avoid oversimplifying and reducing the philosophical complexity and cultural richness of ethical principles and concepts which should in-

form its design; in parallel, we should expand and revise the understanding of such ethical principles and the related values by considering also non-mainstream and non-Western ethical accounts. In this last section, we provide a few insights to show briefly an example of such an operation and highlight how considering relational theories developed in both Western and non-Western moral philosophy (see, for example, feminist ethics and African philosophy of Ubuntu) can help to decolonize AI and AI ethics principles, namely: to center the most marginalized in the design of AI, considering the social conditions of oppression often tied to the legacy of colonization affecting them, and make AI systems tools truly enabling to mitigate existing systemic unfair inequalities.

Let us do this by continuing our previous example on autonomy. To avoid oversimplifying ethics, a proper ethical inquiry on autonomy drawing insights on moral philosophy would show that autonomy and rational self-determination do not overlap, and that autonomy encompasses a relational dimension too. Such a relational understanding emerges both in non-mainstream Western approaches that are currently at the outskirt of the debate on AI ethics and AI design (e.g., communitarianism, feminist ethics, etc.), as well as in non-ethnocentric ethical accounts (consider the African philosophy of Ubuntu ethics). Therefore, it sounds precious to consider such theories to revise the current mainly liberal notion of autonomy as self-government and independence underpinning the AI ethics discourse and AI design (we can operate similarly for other AI ethics principles, such as fairness). We will not expand such relational accounts here, due to space constraints. However, highlighting some of their criticism to the mainstream or standard notion of autonomy sounds to be precious to understand the need of using AI to mitigate unfair inequalities. For example, relational scholars highlight the importance of not focusing exclusively on rational deliberation, insofar as very often the options on which we choose or our reflective judgments are already tainted by oppression, which makes our autonomous choices just an act of rationalization of oppressive concepts and biased ways of thinking and living. Despite their heterogenous views, relational scholars tend to widely agree on investigating social and relational conditions and contexts in which we live to detect and examine such sources of oppression, as well as in finding in socio-relational conditions and aspects some key requirements (e.g., social support and recognition in Communitarianism, and/or solidarity in Ubuntu ethics) to both enable and empower the agency of individuals already undermined by oppression and inequalities of various kind.

From considering such relational ethical views we can point out a few ethically meaningful implications for the design of AI to contribute to a fairer society. Indeed, according to such relational views, for example, we

should avoid applying mathematical neutral or parity models to design fairer AI. Such models indeed lead us to just shallowly fix bias by leveling differences in AI performance and outputs: they do not change or mitigate, but instead legitimate, real-world inequalities of people in accessing or benefitting from such systems. We should instead use AI to discover and investigate why such unfair bias emerges in certain relational and social contexts, what historically asymmetrical relations of power they reflect, and what and who continue to nurture them – information very often invisible to us and that instead AI systems thanks to their capacity to infer patterns and correlations from processing huge amounts of data can unveil.

Without the pretense of unpacking all the key implications such relational theories developed in various contexts can bring out for the design of AI, we hope such article have shown their value and will spur further research on how such approaches, along with other multicultural philosophical perspectives, will be more and more crucial to ethically design AI systems that can help us to mitigate unfair inequalities and injustice in our globalized world.

Conclusion

In this paper we have addressed one of the most discussed risks in the field of AI ethics, that is, the risk of AI perpetuating and exacerbating existing and historical inequalities. More specifically, we have shown how current AI ethics principles might be inadequate to design and develop AI as a tool promoting fairer and more just societies, unless they are properly understood and decolonized. To this aim, after having clarified the issue of inequalities in relation to AI, we have criticized the current main ethical approach to AI to be oversimplified and limited, as reflecting just mainstream Western ethical theory. Specifically, we have argued how such an ethical approach is at odds with the design of a fairer and more inclusive AI, as it fails to properly consider many forms of oppression affecting the conditions in which many live. To overcome such limits, we proposed to revise current AI ethics principles with relational theories, developed in Western and non-Western moral philosophy, insofar as they allow us to more properly consider the many and different social and cultural aspects shaping individuals and the conditions in which they express their agency, including historical and systemic forms of oppression and injustice affecting them, and therefore, to adequately design AI capable to effectively mitigate or prevent existing unfair inequalities and empower the most vulnerable, marginalized, and thus far most negatively affected by AI.

Bibliography

Angwin J., et al.
2016 *Machine bias*, in "ProPublica", available at:
https:// www.propublica.org/article/machine-bias-risk-assessments-in- criminal-sentencing.

Bietti, E.,
2019 *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*, in "Proceedings to ACM FAT* Conference (FAT* 2020)", Available at: https://ssrn.com/abstract=3513182

Binns, R.
2018 *Fairness in machine learning: lessons from political philosophy*, in "Arxiv", available at:
http://arxiv.org/abs/1712.03586.

Calvo, R., et al.
2020 *Supporting human autonomy in AI systems: a framework for ethical enquiry*, in Burr, C., Floridi. L. (eds.), *Ethics of digital well-being: philosophical studies series*, Springer, Cham.

Couldry, N., Mejias, U.
2019 *The costs of connection: how data colonizes human life and appropriates it for capitalism*, Stanford University Press, Stanford.

Dignum, V.
2022 *Relational artificial intelligence*, in "Arxiv". Available at: https://doi.org/10. 48550/arXiv.2202.07446.

European Parliament
2017 *Report with recommendations to the Commission on Civil Law Rules on Robotics*. Available at: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html

Floridi, L. et al.
2018 *AI4People–an ethical framework for a good AI society: opportunities, risk, principles, and recommendations*, in "Minds Machines", 28, pp. 689-707.

Floridi, L., Cowls, J.
2019 *A unified framework of five principles for AI in society*, in "Harvard Data Science Review", 1,1.

Fjeld, J. *et al.*
2020 *Principled Artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI*, in "Berkman Klein Center Research Publication", 1.

Fricker, M.
2007 *Epistemic injustice: power and the ethics of knowing*. Oxford University Press, New York.

Giovanola, B., Tiribelli, S.
2022 *Weapons of Moral construction? On the value of fairness in algorithmic decision-making*, in "Ethics and Information Technology", 24, 1-13.

Jaworska, A.
2009 *Caring, minimal autonomy, and the limits of liberalism*, in Lindemann, H., Verkerk, M., Walker, M. (eds.), *Naturalized bioethics: toward responsible knowing and practice*, Cambridge University Press, Cambridge.

Jobin A., et al.
2019 *Artificial intelligence: the global landscape of ethics guidelines*, in "Nature Machine Intelligence", 1, pp. 389-399.

Kahneman, D.
2011 Thinking fast and slow, Straus & Giroux, New York.

Mhlambi, S., Tiribelli, S.
2023 *Decolonizing AI Ethics: Relational autonomy as a means to counter AI harms*, in "Topoi", pp. 1-14.

Mittelstadt, B.D., et al.
2016 *The Ethics of Algorithms: Mapping the Debate*, in "Big Data & Society", 3, 2, pp. 1-21.

Mittelstadt, B., Watcher, S., Russell, C.
2023 *The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default*, Available at: https://ssrn.com/abstract=4331652

Milano, S., Taddeo, M., Floridi, L.
2020 *Recommender systems and their ethical challenges*, AI & Society, 35, pp. 957-967.

Mohamed, S., Ping, M.T., Isaac, W.
2020 Decolonial AI: *decolonial theory as sociotechnical foresight in artificial intelligence*, in "Philosophy and Technology", 33, pp. 659-684.

Obermeyer Z, et al.
2019 *Dissecting racial bias in an algorithm used to manage the health of populations*, in "Science" 366, pp. 447-453.

Pasquale, F.
2015 *The Black Box Society: The Secret Algorithms that Control Money and Information*, Harvard University Press, Cambridge.

Prunkl, C.
2022 *Human autonomy in the age of artificial intelligence*, in "Nature Machine Intelligence", 4, pp. 99-101.

Rességuier, A., Rodrigues, R.
2020 *AI ethics should not remain toothless! A call to bring back the teeth of ethics*, in "Big Data & Society", 7, 2.

Simon, H.
1991 *Bounded rationality and organizational learning*, in "Organizational Science", 2,1, pp. 125-134.

Selbst, A.
2019 *Fairness and abstraction in sociotechnical systems*, in "Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19, 59-68", ACM Press, Atlanta, GA, USA.
Simon 1991

Thaler, R., Sunstein, C.
2009 *Nudge: Improving decisions about health, wealth and happiness*, Penguin, London.

Tiribelli, S.
2023 *The AI ethics principle of autonomy in health recommender systems*, in "Argumenta", 16, pp. 1-18.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.
2022 *The Ethics of Algorithms: Key Problems and Solutions*, in "AI & Society", 37, pp. 215-230.

World Health Organization
2021 *Ethics and Governance of Artificial Intelligence for Health.* Available at: https://www.who.int/publications/i/item/ 9789240029200

Zuboff, S.
2019 *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Public Affairs, New York.