AI Ethics for Sustainable Development Goals

Aníbal Monasterio Astobiza
Universidad de Granada, 18011 Granada, Spain

Mario Toboso
IFS-CSIC (Spanish National Research Council), 28037 Madrid, Spain

Manuel Aparicio
University of Murcia, 30003 Murcia, Spain

Daniel López
IFS-CSIC (Spanish National Research Council), 28037 Madrid, Spain

■ **We live in** an era where problems are global in scale (e.g., climate change) and solutions must also be coordinated on a global scale in the international context. The 2030 agenda for sustainable development goals (SDGs) was ratified in 2015 as a continuation of the millennium development goals (MDGs). In this sense, the SDGs, as MDGs were in their day, are a global mechanism that urges governments to coordinate to address global problems. At the core of the SDGs is "to achieve a better and more sustainable future for all." The SDGs consist in a series of 17 goals with 169 targets which, for the first time, identify the fight against poverty as a necessity for sustainable development. The SDGs consider the ecological, social, and economic dimensions as interdependent for sustainable development. With the progress and advances of artificial intelligence (AI) technologies, many researchers are exploring the possibility of their use to tackle societal problems. This is what many people nowadays call "AI for social good" (AI4SG). The concept behind AI4SG is very simple: AI-powered systems and capabilities applied to improve public welfare [1]. Although there are different forms of classification of AI4SG initiatives (in terms of data, modeling, or decision-making) "projects addressing AI4SG vary significantly" [2] and the AI behind these projects may have been designed for the "good" but, in practice, it could end up going "bad." More importantly, not everyone would agree on what is a good result. The main motivation for any application of the AI4SG is to solve social problems.

Digital Object Identifier 10.1109/MTS.2021.3056294
Date of current version: 3 June 2021.

However, the development of an algorithm in a laboratory without the possibility of real-world application is not AI4SG. The lack of AI research for proof of concept or pilot testing with real-world application and implementation often makes it very difficult to evaluate many AI4SG projects. On the other hand, at the background, there may be sincere motivation and good intentions, but the application of AI may create other social problems in addition to those initially intended to be solved. For example, cities around the world strive hard to reduce antisocial and violent behavior, leading the city authorities to rely on the use of AI systems that sketch the profile of the population, that is, classify and predict people's antisocial behavior. However, this has often resulted in discrimination against certain groups. Algorithms designed to predict antisocial behavior were fed with not very robust data from statistics of illicit acts leading to surveillance, and the simple mass surveillance of these neighborhoods led to more violations being reported for strictly statistical reasons which in turn served to continue to feed the algorithm with the consequent effect of self-fulfilling prophecy: minority ethnic groups and stigmatized groups are discriminated because of prejudice [3]. Preventing this from happening is in the realm of ethics that must be taken into account from the start to the end, and engineers and computer scientists that want to deploy AI technologies to solve complex problems, first of all, bear ethics in mind while designing the AI technologies. Obviously, not always everyone knows what is involved in ethical design.

0278-0097/21©2021 IEEE

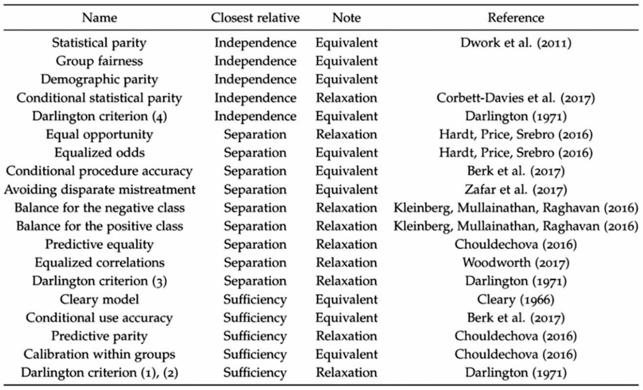
What is “ethics” in AI ethics?

Even though the term “ethics” cannot have a straightforward definition, nor can it entertain anyone who wishes to define, philosophers differentiate between “morality” and “ethics.” Morality is descriptive, in the sense that covers the actual behavior of individuals, while ethics is prescriptive in nature and dictates what individuals should do. In other words, morality answers the question: what do people do? and ethics answers the question: what should people do? Many AI researchers and practitioners have never heard of this distinction, nor even take any “ethics 101” class or lesson, even though they build AI-powered systems or socio-technological systems that interact in the world and can have an impact on individuals and society as a whole. What is understood by ethics in the computer sciences curricula is not the same as what is understood by ethics in philosophy, political philosophy, and other social and human sciences. To put it bluntly, ethics is an area of study with a long tradition (spans several millennia) that deals with complex concepts and actions that draw on the insights of multiple disciplines that have emerged throughout history such as anthropology, economics, psychology, etc., to name a few. In other words, ethics is hard. Let us quote in length a recent treatment in book format about how to decide among different ethical theories in the face of uncertainty to grasp how hard ethics is: “As with other areas of philosophy, working out the correct moral view often involves being sensitive to subtle distinctions, being able to hold in mind many different arguments for different views, and paying attention to intuitions across many different thought experiments. It also involves difficult questions about how to weigh different theoretical virtues, such as simplicity and elegance against intuitive plausibility. Correctly balancing all these different considerations is extremely difficult, so even when we come to a firm stance about some ethical view, we should not always expect that our reasoning is error-free” [4, p. 11]. “Ethics” in AI ethics for the promotion of SDGs or any other social good aims to investigate and identify critical ethical issues when building AI capabilities, and the main concern of ethics is to identify product vulnerabilities and recommend strategies for prevention based on sound ethical analysis. Laypeople and people outside the professional field of philosophy usually use the terms ethics, morality, and even law interchangeably. As we have shown above, ethics and morality are two distinct concepts. To be more specific, morality is a set of norms, values, principles, etc., to behave correctly which are shared by a group. On the other hand, ethics is the philosophical study of norms, values, principles, etc., and what it means to behave correctly or wrongfully. And law is a set of formal rules which confers duties and rights to organize society. What is “ethics” in AI ethics is a question of value alignment as well [5]. The main problem with the alignment of values or ethically aligned AI is how we decide which values to implement given that there is no agreed definition of what ethics is and of course we live in pluralistic societies where people differ in which values are most important. But there is also a problem with what we have to understand by AI. AI can be understood as the science and engineering that seeks to create intelligent machines. For the purpose of this article, we understand by AI any artificial system that seeks to fulfill its objectives, whatever these may be. AI value alignment refers to how we make an artificial system compliant with human values [6]. An important aspect to consider in the AI value alignment issue is the type of method we use to build an AI system and the encoding of values in those systems: the technical and the axiological. Within machine learning, a set of techniques and methods that allow machines to learn from data to make predictions and improve decision-making to create intelligent machines, there are different subfields and each of these different subfields constrains the type of values or ethical principles to be codified in the design of an AI model.

- *Supervised learning (SL):* Train an AI model to perform a task with labeled data.
- *Unsupervised learning (UL):* Train an AI model to perform a task with unlabeled data.
- *Reinforcement learning (RL):* An agent learns to maximize a reward signal from the environment.

Each of these technical methods for building an AI seems to correspond to an established classical ethical doctrine [5]. For example, SL is *Kantian* or deontological in its characteristics. SL follows the Kantian maxim: “Act only on that maxim through which you can at the same time will that it should become a universal law [of nature]” but in this case applied to an AI system. In other words, the programmer or the engineer labels the data to train the AI model and therefore codes some values or principles considered by him as reasonable and universal that have to guide the system’s behavior. Meanwhile, UL seems to follow an Aristotelian virtue ethics in the sense that it allows the AI algorithm to find patterns in the unlabeled data for a solution to the required task. In other words, because this kind of ethical theory, Aristotelian virtue ethics, is built on self-realization and UL trying to uncover patterns in unlabeled data is like as if the system self-realizes to achieve the required task. For the case of RL, the parallelism with another classical ethical doctrine, utilitarianism, is clear. An agent (AI system) wants to maximize the reward signal from the environment and according to act utilitarianism a right action is the one that creates the greatest happiness for the greatest number of people. The tradeoff between these two aspects, the axiological and technological, is not always easy because they are intertwined. So, the type of AI technique chose influences the principles and values that will constrain the behavior of the system. But here we enter into another fundamental problem of a philosophical nature. The distinction between facts and values [7]. Many of the techniques within machine learning are trained with data (facts) but these data do not have to contain correct "values." That is, even if an AI model imitates what a programmer has labeled as correct or discovers patterns in the data, these newly discovered patterns do not have to be axiologically correct or just because they imitate what a human being has done they have to be correct or valuable. And this problem is exacerbated when you decide to design an AI technology the ultimate goal of which is to do social good such as the promotion of SDGs.

Table 1. List of some specific metrics of AI research on fairness. Source taken from *Fairness and Machine Learning* by Solon Barocas, Moritz Hardt, and Arvind Narayanan.



Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Ethics by design

IEEE's ethically aligned design (https://ethicsinaction.ieee.org/) puts a strong emphasis in human wellbeing when building AI technologies. Montréal declaration for a responsible development of AI (https://www.montrealdeclaration-responsibleai.com/the-declaration) defends certain values and ethical principles that protect individuals and groups. As AI technologies become more prevalent in society, it becomes more than necessary to implement principles of regulation and control of AI. But the number of guidelines, manifestos, statements by industry, governments, and civil society is growing rapidly [8]. And none is strong enough to be used widely by all stakeholders. Most of these guidelines are analytically relevant, accurate, but often have a simplified view of ethics as discussed above. As we will present in the "HR as a normative framework for AI technologies" section, the best normative guide to establish ethical principles of governance of AI technologies is the human rights (HR) framework. But first, we would like to point out some basic principles that some of these guidelines discuss and that are necessary for an AI ethics.

Fairness

Fairness is central to human moral cognition but at the same time is very controversial too. Its definition, its origins, or the possible presence in other species is elusive and a source of conflict between a large number of disciplines. If we correctly understand what fairness means our society can be better organized [9]. Fairness is not the same as equality. Equality means same distribution of resources among all, while fairness means just distribution of resources among all. The following example clarifies the difference. An equal distribution of four apples between two people is that each gets two. But a fair distribution is one when one of the two has contributed more, say, has planted the tree and cared for it and picked the apples, he takes more apples for the effort in contrast to the other who has done nothing. In this sense, fairness is equal to justice or just causes. In relation to the use of an AI system, in order for it to be fair does not have to produce equal outcomes. Many specific metrics in AI research on fairness focus on a parity outcome (see Table 1) and this is not fairness properly speaking. A recent article [10] has reviewed the three main mathematical definitions of fairness concluding that all of them suffer from significant statistical limitations. This is a clear example of the use and misuse of fairness in machine learning research. Consequently, methods for auditing algorithms used in machine learning research applied to medicine, transport, banking, finance, or even as an AI4SG initiative have to be ethically informed *ex ante, durante actione*, and *ex post* to avoid building definitions of fairness from what the needs of computer science are instead of true and contextual fairness.

Transparency

Many ethical guidelines overestimate the principle of transparency. They consider that many AI models are very complex and often act as black boxes, say, it is not very well known how machine learning models obtain results. The real issue here is that people are not much more transparent than AI systems and that does not stop us from trusting people. Nonetheless, transparency is important insofar is related to the issue of explainability and when it comes to AI systems applied for SDGs, it is necessary to have an explanation of their operation in case they end up doing something wrong.

Explainability

At the European level, the first regulations banning the possibility to make certain decisions based on algorithms are beginning to arrive [11]. Citizens now have the right to be told how and why an algorithm has made a certain decision: a right to explanation. So, in this sense, explainability is the degree in which a human understands a machine model. For an ethically correct explanation of an artificial system, there must be the possibility of generation of explanations for algorithmic outputs but it is also necessary to have tools and methodologies for conducting algorithm audits. Because one of the biggest problems are that machine learning or deep learning algorithms sometimes fail, and we do not know why.

Global governance of algorithmic systems

Once you have an AI model ready to be used to meet the SDGs and is fair, transparent, and explainable, you must bear in mind that the governance of AI technologies must be global. The international competition for leadership in the digital revolution tends to focus on acquiring new skills, much less on the governance, control, and regulation of the social and ethical impact of emerging AI technologies. When we talk about collective interests or global governance of AI, we see that achieving milestones is not a zero-sum game, because nothing is lost by sharing new capabilities, algorithms, human capital, infrastructures, etc. This is why the global governance of AI must benefit all human beings, including future generations. It is, therefore, necessary to ensure that AI technologies are sustainable and environmentally friendly. Furthermore, they must take into account the environment [12], including other living beings, and their social impact must be carefully considered. Accountability is another major factor to take into consideration. Mechanisms should be put in place to ensure the responsibility and accountability of AI systems and their results. Auditing, which allows the evaluation of algorithms, data, and design processes, plays a key role in this, especially in critical applications. In addition, access and redress to those affected by algorithmic decisions must be guaranteed. History reveals that there are many forms of governance, but AI and robotics pose a challenge to existing governance models. It is important to understand that because regulations exist, the world is safer and more reliable. Regulations are instruments used to implement social or policy objectives. In some sense, these instruments encourage just and fair outcomes and are implemented to correct externalities and other failures. With the appropriate AI governance model based on the HR framework, one can determine how to use AI and robotics to advance SDGs.

HR as a normative framework for AI technologies

Despite the large number of guidelines and principle-based approaches to AI ethics, the only theoretical and practical framework on which to base a global governance of AI and, in particular, to its application for the SDGs is HR. HR should constrain any design or deployment of AI technologies. Any use of AI must respect people's individual and civil rights. Any belief that technology is neutral must be eliminated. The supposed neutrality of algorithms leads to hiding the asymmetries of power between the ruling class and oppressed minorities. The use of supposedly neutral and entirely rational technology can be used as an excuse to maintain unjust social hierarchies. The way in which AI is used has implications that can disenfranchise or empower humanity. If the use of AI is made on the basis of taking into consideration HR respect, and defense of individual experience will always be the main guide. But there is also a criticism of the modern, enlightened conception of the values that permeate HR. HR puts too much emphasis on the ethics of personhood or individual. The philosophy of ubu-Ntu through its robust conception of relationality invites us to reimagine not only digital technologies, the Internet, and AI but what it means to be human in an interconnected relationship with the natural world and other forms of life [13]. In the age of AI and robots, the HR must take the central stage. Individual HR such as the integrity of the person, autonomy, self-determination, the right to privacy but also the right to nondiscrimination, free speech, and the right to political participation should guide the development and implementation of AI technologies. If directly or indirectly these HR are compromised in the design or implementation of an AI technology, this should cancel their deployment.

Human-centric AI and SDGs

Modern economy and society has evolved by adopting technical advances such as machines for automation of many tasks and activities. This automation and deployment of machines including AI technologies have helped society to try meet the SDGs in a predictable manner. Society needs smart and precise solutions and these are becoming available through new applications of AI technologies. Some examples of these applications of AI technologies to meet the SDGs are

- sensors to measure CO_2 emissions;
- smart farming: machine learning applications to improve horticultural products;
- infection disease tracking and monitoring system; and
- GPS and satellite imagery to monitor wildlife.

However, whatever applications are used to meet the SDGs must be made from an ethical and humane-centric perspective. With respect to the ethical part, as we have commented above, this must be done from a rigorous approach to ethics and not from a simplified vision of it. Regarding the humane dimension, every AI technology must be applied in accordance to the HR framework. One of the main axioms of human-centric AI is to do everything possible to achieve value alignment: how to align the values of highly capable intelligent machines with those of humanity. Value alignment problem consists in how to align AI with goals, values, and preferences of its users when the potential users can be all of humanity. A practical dimension of the value alignment problem is to create artificial systems that meet people's needs. AI applied to SDGs is the corollary of the value alignment problem. If we were able to harness the power of AI to deliver on the SDGs, the problem of value alignment could be largely solved and we will have created a human-centered AI. Human-centric AI means developing a new generation of AI models that perform operations and tasks aimed at improving the welfare of all, or at least, caring for the planet, communities, and individuals.

Discussion

Ethics is an age-old discipline that should be taken seriously when reflecting on the impact of AI technologies. Most mathematical approaches to several concepts within AI ethics (e.g., fairness) do not take into account the complexity of what "ethics" implies. Here, we wanted to emphasize the idea that a principle-based approach to AI ethics is not enough and that we must take HR as a basic normative framework for AI technologies. Our take home message is that to meet the SDGs with the help of AI technologies, we must consider ethics from the start (in the designing and building of AI technologies). Another aspect to bear in mind when dealing with AI ethics applied to SDGS is acknowledging the mutual reinforcement between three different components involved in any algorithmic system:

$$\begin{array}{c}\text{Algorithms}\\ \updownarrow\\ \text{People}\\ \updownarrow\\ \text{Data}\end{array}$$

In this sense, the outcomes of an AI system will depend on the mathematical and technical aspects of the model, the data with which the model is trained and the people who interact with the algorithmic system. At any of these points in the process there may be unintended consequences (e.g., bias) and with a much greater impact, inherent complexity or dimensionality when it comes to the deployment of AI-powered systems in the context of SDGs, so a rigorous ethical analysis of each of these components is a necessity. One recommendation is to hire a professional philosopher or an ethicist for your AI project or initiative to support and advance the SDGs. On the other hand, the global governance of AI requires international cooperation between countries because using AI technologies to meet the SDGs requires coordination beyond nation states or individual countries. To this end, supranational bodies such as the UN have to lead the global governance of AI. ■

Acknowledgments

This work was supported in part by the European projects INBOTS under Grant 780073 and in part by the EXTEND of the H2020 Programme under Grant 779982 and in part by EthAI+3 under Grant PID2019-104943RB-100. We would also like to thank the organizers of the workshop on "AI4EQ: Advancing Towards the SDGS: Artificial Intelligence for a Fair, Just, and Equitable World," ECAI2020, Santiago de Compostela, for their time and effort. We are also indebted to the reviewers for their comments.

References

[1] G. D. Hager et al., "Artificial intelligence for social good," in *Proc. CCC Workshop Rep.*, 2017, p. 24.

[2] J. Cowls et al., *Designing AI for social good: Seven essential factors*. Accessed: Feb. 2020. [Online]. Available: https://ssrn.com/abstract=3388669

[3] R. Richardson, J. M. Schultz, and K. Crawford, "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice," *NYUL Rev. Online*, vol. 94, p. 15, Mar. 2019. [Online]. Available: https://ssrn.com/abstract=3333423

[4] W. Macaskill, K. Bykvist, and T. Ord, *Moral Uncertainty*. Oxford, U.K.: Oxford Univ. Press, 2020.

[5] I. Gabriel, "Artificial intelligence, values and alignment," 2020, *arXiv:2001.09768*. [Online]. Available: http://arxiv.org/abs/2001.09768

[6] S. Russell, *Human Compatible: AI and the Problem of Control*. New York, NY, USA: Allen Lane, 2019.

[7] H. Putnam, *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA, USA: Harvard Univ. Press, 2004.

[8] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[9] J. Rawls, *Justice as Fairness: A Restatement*. Cambridge, MA, USA: Harvard Univ. Press, 2001.

[10] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," 2018, *arXiv:1808.00023*. [Online]. Available: http://arxiv.org/abs/1808.00023

[11] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.

[12] R. Schwartz et al., "Green AI," 2019, *arXiv:1907.10597*. [Online]. Available: http://arxiv.org/abs/1907.10597

[13] S. Mhlambi, "From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance," Carr Center Discussion Paper series 2020-009, 2020.

Aníbal Monasterio Astobiza is currently a postdoctoral researcher at Departamento de Filosofía I, Universidad de Granada, Granada, Spain.

He is a member of the INBOTS Task Force (Horizon 2020, CSA, ref. 780073), the EXTEND Task Force (Horizon 2020 Research Project, ref. 779982), IFS-CSIC, and the EthAI+3 (Digital Ethics: Moral Enhancement Through an Interactive Use of Artificial Intelligence, PID2019-104943RB-100).

Mario Toboso is a Tenured Scientist with IFS-CSIC, Madrid, Spain.

He is a member of the INBOTS Task Force (Horizon 2020, CSA, ref. 780073) and the EXTEND Task Force (Horizon 2020 Research Project, ref. 779982), IFS-CSIC.

Manuel Aparicio is an Assistant Professor with the University of Murcia, Murcia, Spain.

He is a member of the INBOTS Task Force (Horizon 2020, CSA, ref. 780073) and the EXTEND Task Force (Horizon 2020 Research Project, ref. 779982), IFS-CSIC.

Daniel López is a Ph.D. Student with IFS-CSIC, Madrid, Spain.

He is a member of the INBOTS Task Force (Horizon 2020, CSA, ref. 780073) and the EXTEND Task Force (Horizon 2020 Research Project, ref. 779982), IFS-CSIC.

■ Direct questions and comments about this article to Aníbal Monasterio Astobiza, Departamento de Filosofía I, Universidad de Granada, 18011 Granada, Spain; anibalmastobiza@gmail.com