

Ethical assessments and mitigation strategies for biases in AI-systems used during the COVID-19 pandemic

Big Data & Society
January–June: 1–11
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517231179199
journals.sagepub.com/home/bds



Alicia de Manuel¹ , Janet Delgado² , Iris Parra Jounou¹,
Txetxu Ausín³ , David Casacuberta¹, Maite Cruz⁴,
Ariel Guersenzvaig⁵, Cristian Moyano¹, David Rodríguez-Arias²,
Jon Rueda⁶ and Angel Puyol¹

Abstract

The main aim of this article is to reflect on the impact of biases related to artificial intelligence (AI) systems developed to tackle issues arising from the COVID-19 pandemic, with special focus on those developed for triage and risk prediction. A secondary aim is to review assessment tools that have been developed to prevent biases in AI systems. In addition, we provide a conceptual clarification for some terms related to biases in this particular context. We focus mainly on non-racial biases that may be less considered when addressing biases in AI systems in the existing literature. In the manuscript, we found that the existence of bias in AI systems used for COVID-19 can result in algorithmic justice and that the legal frameworks and strategies developed to prevent the apparition of bias have failed to adequately consider social determinants of health. Finally, we make some recommendations on how to include more diverse professional profiles in order to develop AI systems that increase the epistemic diversity needed to tackle AI biases during the COVID-19 pandemic and beyond.

Keywords

AI systems, bias, triage and risk prediction, social determinants of health, COVID-19

Background

Pandemics such as COVID-19 are eminently complex to process in the healthcare field, both at an organizational and a cognitive level. For this reason, some researchers have been developing automatic algorithms that facilitate decision-making for healthcare staff. While these algorithms must be functional and based on as much complete data as possible, they must also be free from biases or prejudices. Automating an algorithm may multiply the harmful effects of any bias in its design and application. This downside can be even more frequent in algorithms that are urgently developed and implemented during a pandemic. Thus, even assuming data reliability, accuracy and veracity, it is still worth questioning whether the automatic decision-making system is a just one (Pot et al., 2021). In this respect, it is both an ethical duty and a quality requirement to ensure that the system respects algorithmic justice, which implies addressing and taking responsibility for disputes and harm caused by automated algorithm decision-making that go beyond issues of social justice (Marjanovic et al., 2021).

The main aim of this article is to reflect on the impact of biases related to artificial intelligence (AI) systems developed to tackle issues arising from the COVID-19 pandemic, with special focus on those developed for triage and risk prediction. A secondary aim is to review assessment tools

¹Department of Philosophy, Universitat Autònoma de Barcelona, Barcelona, Spain

²Department of Philosophy, Faculty of Philosophy, University of Granada, Granada, Spain

³Institute of Philosophy of the CSIC (Spanish National Research Council), Madrid, Spain

⁴Andalusian School of Public Health (EASP), Granada, Spain

⁵ELISAVA Barcelona School of Design and Engineering UVIC-UCC, Barcelona, Spain

⁶FiloLab Scientific Unit of Excellence, University of Granada, Granada, Spain

Corresponding author:

Alicia de Manuel, Department of Philosophy, Universitat Autònoma de Barcelona, Barcelona, Spain.

Email: aliciademanellozano@gmail.com



that have been developed to prevent biases in AI systems. In addition, we provide a conceptual clarification for some terms related to biases in this particular context. Due to the amount of literature addressing racial biases in AI systems before the pandemic (Kostick-Quenet et al., 2022; Livingston, 2020; Noor, 2020; Noseworthy et al., 2020; Tat et al., 2020; Turner Lee, 2018), as well as during it (Leslie et al., 2021; Luengo-Oroz et al., 2021; Rööslé et al., 2021; Williams et al., 2020), we focus mainly on non-racial biases that may be less considered when addressing biases in AI systems in the existing literature. Nevertheless, from an intersectional approach, racial biases are profoundly connected to other biases caused by an overlook of Social Determinants of Health (SDOH) and remain of major importance. It is important to study the impact of SDOH in AI biases to achieve a better understanding of racial biases as well.

We will start by clarifying key terms related to biases in the context of AI systems developed to tackle COVID-19 issues. The terms we will explain are as follows: bias; AI algorithms; triage and risk prediction; algorithmic justice; and SDOH. Secondly, we will explore some of the ethical issues that have been neglected in AI systems developed in the context of the COVID-19 pandemic. And finally, we will comment on the different assessment tools and regulations that are being devised to prevent the appearance of biases in AI systems.

Definitions and contextualization

Due to the ambiguity or vagueness of some terms, we provide here a list of definitions to clarify the meaning of those we are using in the context of this study (Table 1).

Neglected ethical problems

During the COVID-19 crisis, the development of AI has increased due to its capacity to improve data management. It has been mainly designed to help allocate healthcare resources through diagnosis and patient risk prediction, and monitor the evolution of the pandemic and control population spread (Council of Europe, no date).

In previous research (Delgado et al., 2022), we found that AI systems have been mainly employed in triage and patient risk prediction and CTApps. Even though the implementation of AI has offered many benefits, the huge amount of data involved and the rapid rate of technological implementation have generated important ethical issues related to the appearance of biases in these areas of implementation.

Epidemiologically effective, but unethical

The COVID-19 pandemic has shown that systems based on machine learning (ML) can benefit the health of a group (Gao et al., 2020; Quiroz-Juárez et al., 2021). Such AI

applications have made it possible to efficiently diagnose people at risk of the disease and predict its evolution. ML models have demonstrated epidemiological effectiveness in controlling some of the public health impacts of the pandemic. However, the collective health benefits that are generated do not necessarily justify the morality of ML. In general terms, despite the possible epidemiological advantages, if the application of ML-based systems leads to certain ethical biases that discriminate against individuals or groups for morally arbitrary reasons such as gender, race, culture and socio-economic status, then this may be deemed immoral.

For example, let us consider a Patient Risk Prediction App based on a disability-biased ML to allocate scarce resources as ventilators. The system succeeds in medical rationing according to ableist parameters like life expectancy (long-term survival, short-term survival and reasonable accommodation) and quality of life (Goggin and Ellis, 2020) but disabled people are systematically denied a ventilator. The public health aim has been achieved, but the measure is immoral. In such a case, there is a conflict between public health goals and social justice. The existence of these ethical conflicts needs to always be borne in mind.

To determine the morality of an epidemiological action (which includes the creation of algorithms that pursue greater epidemiological effectiveness), it is not enough to focus only on their effectiveness, but also on the morality of the entire process from the origin of the action to the result. The public good embodied in public health measures is not only defined by the fact that the entire community benefits from them, but also that such measures are not based on ML systems and algorithms that contain unfairly discriminatory biases against certain individuals or social groups.

What place (if any) do SDOH have in AI systems?

Our previous analysis of studies on AI systems developed during the COVID-19 pandemic (Delgado et al., 2022) revealed that there is no systematic concern for parameters related to SDOH during the processes of data collection, design, and implementation of these AI systems in health-care. At the same time, no relevant studies were found on the relationship between algorithmic injustice and SDOH for such cases. This overlook can also be seen in other relevant aspects like age, as Chu et al. showed (2022).

As social conditions that can affect health risks and outcomes, SDOH are keys to understanding unfair health inequities based on systematic discrimination. We focus mainly on non-racial biases that may be less considered when addressing biases in AI systems. Nevertheless, from an intersectional framework, racial biases remain of crucial importance and are profoundly connected to other biases (disability, age, gender, etc.) related to SDOH.

Table 1. Definitions, clarifications and contextualization.

Bias	
Definition	“Strong inclination either in favor or against something” (Moseley, 2021). In relation to AI, algorithmic biases are systematic errors in a computer system, which means a deviation from the expected prediction behavior of an AI tool (Amann et al., 2020). A bias can arise from the algorithm design and/or the previous data collection, coding and selection of data (technical and computational matters) or from inappropriate uses or deployment of the algorithm (Danks and London, 2017). Such biases may generate results that are systemically prejudiced against a segment of the population due to erroneous assumptions in the machine-learning process or due to its use in a population for which it was not designed.
Types	According to Danks and London’s taxonomy (2017), biases can be classified into training data bias (deviations in the training or input data; creating the algorithm from previous databases can generate unfair results if the data is not representative of population diversity, or some segment is overrepresented while others are underrepresented), algorithmic focus bias (non-use of information, for example, not to include older adults in the ICU based on their low chances of survival), algorithmic processing bias (statistically biased estimator), transfer context bias (use outside the context it was designed for) and interpretation bias (misinterpretation of the outcome by the user).
Consequences for healthcare related to the COVID-19 pandemic	In healthcare, if there are biases in the input data, then there could be prediction errors as some population groups could be underrepresented in the training sample (Amann et al., 2020), which can lead to unfair results such as privileging one group of users over another. Although algorithms are commonly presented as neutral mathematical models, the fact is that they are instructions that can reproduce ethically undesirable arbitrariness and discrimination (Cossette-Lefebvre and Maclure, 2022).
Examples	Chorás et al. (2020) identified potential causes of biases in AI: (a) skewed sample: misrepresentation of training data in some areas that evolves over time; (b) tainted examples: existing bias because of old data caused by human bias can be replicated by the system; (c) limited features: the system can cause lower precision in predictions for minority groups because of unreliable or less informative data collection; (d) sample size disparity: if there is an insufficient data sample, this can cause difficulties in building a reliable model for the group; (e) proxies: the expected output can have some bias due to correlations of biased sensitive attributes with other characteristics.
Machine-Learning Algorithm	
Definition	An algorithm is described as any set of precise instructions that must be followed in order to carry out a specific task. Algorithms can be used to make calculations, solve problems, or reach decisions. Whereas statistical models use mathematical equations to code information extracted from a dataset, machine-learning algorithms are a specific type of AI algorithm that allow results to be predicted by means of supervised or unsupervised models. These algorithms search for patterns in datasets, and make accurate predictions based on Big Data (Mittelstadt and Floridi, 2016).
Types	There are many types of machine-learning algorithms. Depending on how they work, they can be divided into: supervised learning (among others, decision tree, random forest, linear regression, support vector machine), unsupervised learning (clustering), semi-supervised learning, and reinforcement learning (deep adversarial network).
Consequences for healthcare related to the COVID-19 pandemic	During the COVID-19 pandemic, the design of AI algorithms to aid decision-making in diagnosis, prognosis evaluation, epidemic prediction and drug discovery increased worldwide (Naseem et al., 2020; Wang et al., 2021) due to limited resources and a sense that in order to properly tackle the disease, while health services’ and governments’ response time needed to improve, their decision-making also required standardized criteria. AI algorithms have shown great calculation and implementation capacity in response to the pandemic and the management of medical resources, mostly in high income countries; however, experts have also found that they have resulted in the appearance of a series of risks (Delgado et al., 2022; Korinek and Stiglitz, 2021)
Examples	The lack of transparency in the algorithms can create the so-called <i>AI black-box problem</i> , which occurs when we do not know why an algorithm has reached a certain conclusion. Under these circumstances, the algorithm may hide untraceable biases and create systematic error and injustices in the allocation of health resources.

(continue)

Table 1. continue

Triage and risk prediction	
Definition	<p>Triage is a selection process in which objects are sorted into categories according to criteria defined by certain characteristics. Categories are given an order of importance and priority, with the purpose to maximize benefit and minimize loss in the event of scarce resources (Mezza, 1992).</p> <p>Risk predictions are statistical models that estimate the risk of an event of interest, such as disease incidence. They have potential use in public health and clinical epidemiology (Pfeiffer and Gail, 2011). Although these models were traditionally mathematical, based on discernment and calibration (agreement between observed outcomes and predictions), over recent years they have been developed as AI support systems and machine learning (ML) applications for use in healthcare systems. In a specific situation with limited resources available, there is a great need to target people at highest risk and with the highest expected benefit (Steyerberg et al., 2010).</p>
Functions	<p>Their functions include identifying patients in a life-threatening situation, ensuring prioritization according to classification level, ensuring the reassessment of patients forced to wait, deciding on the most appropriate area to attend to patients and improving patient flow and service congestion, among others (Soler et al., 2010).</p>
Consequences for healthcare related to the COVID-19 pandemic	<p>In the context of healthcare, triage protocols have mainly been developed in emergency settings to increase efficiency in the allocation of healthcare resources. This usually consists of a brief clinical assessment to identify and classify a patient's medical needs and degree of severity. As such, it is a process for managing clinical risk when demand and needs exceed available resources (Soler et al., 2010). Healthcare workers make use of clinical indicators, symptom categories, or vital risk assessments to establish levels of priority in their triage systems according to a system or plan based on an algorithm (Hortal-Carmona et al., 2021).</p>
Distinctions	<p>A distinction should be made between extraordinary situations (e.g., mass casualty disasters, pandemics, etc.) and routine patient management. The difference is the simultaneity of cases, which outstrip resources required for optimal care. Even if any sort of triage implies at least a modest scarcity of healthcare resources, in these cases the medical approach of triage changes and the focus "can no longer be on each individual but must shift to the population as a whole" (Frykberg, 2005). Triage must be distinguished from processes of macro-allocation, such as decisions made by public health policy-makers, legislators or administrators when allocating healthcare funds or other resources to different population groups (Iserson and Moskop, 2007).</p> <p>There is a distinction between predicting a risk and considering how much risk is tolerable (which depends on ethical, political and cultural values). With triage, although it is important to know the statistical result of predictive judgements, its use will ultimately depend on value judgements (e.g., is this risk level acceptable or not?).</p>
Algorithmic injustice	
Definition	<p>Algorithmic injustice arises when the results of automated algorithmic decision-making powered by machine learning lead to unjustified, unfair, and discriminatory outcomes (Grote and Keeling, 2022; Hedden, 2021; Marjanovic et al., 2021).</p>
Impact	<p>Machine-learning algorithms have numerous impacts in terms of which cases are considered alike and how others are treated. This is because, firstly, AI algorithms develop rules from patterns of similarities in the training data that configure their output, and secondly decisions based on these algorithmic-driven systems have an effect on people.</p>
Consequences for healthcare	<p>In the context of healthcare, algorithmic injustice involves all unjustified and avoidable inequities that result from the use of an algorithm in health resource allocation.</p> <p>Compared to other forms of injustice, algorithmic injustice has two specific features: its systematicity (multiplying effect) and its untraceability (hidden biases). However, as a concept, algorithmic justice shares with other forms of justice the challenge of defining what sorts of inequalities are unjustified (e.g., they involve discrimination) and should therefore be avoided. An algorithm creating and multiplying inequality may therefore not be enough for a decision-making mechanism to be unfair, unless a normative – moral or political– standpoint shows why such inequality is unjustified.</p>
Examples	<p>An example of this is when algorithms surreptitiously consider ethically problematic variables such as area of residence when predicting the benefit of a scarce medical resource. If such variables come to influence the output, injustices could arise when following prioritization recommendations partially based on variables that are not strictly medical. Another type of injustice occurs when, while taking into account</p>

(continue)

Table 1. continue

	medically and ethically relevant variables, an algorithm puts underprivileged social groups at a disadvantage or perpetuates unfair health inequalities [see Section 3.4].
Social Determinants of Health	
Definition	Social Determinants of Health (SDOH or SDH) can be defined as the set of personal, social, economic and environmental factors “in which people are born, grow, live, work and age and which determine their health status” (Wilkinson and Marmot, 2003). They are the result of structural drivers such as the distribution of money, power and resources at global, national and local levels, which in turn depend on policy frameworks (Marmot et al., 2012). SDOH are related to both health inequalities and inequity. The former refer to the differences between two people or groups in terms of health outcomes or other relevant factors or characteristics, while the latter is a normative concept. Health inequity refers to inequalities that are unfair or derive from some form of injustice (Pot et al., 2021).
Domains	SDOH are classified around five key domains (Healthy People 2020, no date): Economic Stability, Education, Health and Health Care, Neighborhood and Built Environment and Social and Community Context.
Consequences for healthcare related to the COVID-19 pandemic	The impact of COVID-19 on people and communities related to SDOH has been broadly documented (Abrams and Szeffler, 2020; Paremoer et al., 2021), for instance, Abrams and Szeffler (2020) emphasized the case of homeless families and children living in poverty, who are at higher risk of viral transmission due to crowded living spaces and scarce access to COVID-19 screening and testing facilities. Physical distancing measures, necessary to prevent the spread of COVID-19, are substantially more difficult for those with adverse social determinants, contributing to short-term and long-term morbidity. Children living in poverty who participate in school lunch programs are at a higher risk of food insecurity due to school closures. This can represent a risk to the physical and mental health of these children, including lowering their immune response, and may increase the risk of infectious disease transmission.
Examples	Common SDOH factors that affect health quality are income and social protection, education, unemployment and job insecurity, working life conditions, food insecurity, housing, basic amenities and the environment, early childhood development, social inclusion and non-discrimination, structural conflict or access to affordable health services of decent quality, among others.

Thus, exploring the place of SDOH in AI systems can improve the understanding and mitigation of other biases too. It is of great importance to continue analyzing and connecting the overlapping dimensions that affect the distribution of scarce medical resources in the future.

Even though all disparities are intersectional and part of the core aspects of health, they are seldom considered in either clinical trial developed to design AI and ML support systems or in the *a posteriori* ethical evaluation of these systems. SDOH remain neglected and undervalued in clinical research because the latter still follows a more biology-based conception of health (Afifi et al., 2020; Pasquale, 2021). This bias stems from a prior epistemic problem in medicine, which is a poor awareness of the social aspects of illnesses, an issue inherited by AI. Since AI systems are intended to evaluate the “most at risk” population, in order to reverse disparities and achieve fairness, it is vital that social factors be included in the constitution of these methods (Delgado et al., 2022).

Three keys to understanding the above phenomenon can be summarized as follows:

1. A lack of widespread use of the concept SDOH. Although specific parameters are sometimes highlighted

(such as race or social status), they are not defined as SDOH, which restricts the power of the analysis.

2. Unawareness of SDOH in clinical practice, which results in biases and discrimination in health contexts due to incomplete databases feeding the models.
3. The difficulties of including fairness in ML algorithms and accurate predictive performance in the design and training of AI algorithms (Roy et al., 2021). Since datasets are imbalanced and can create unfair decisions for minority groups, a counterbalance is required in which different inputs can be given different weights in the final prediction. Biases are not normally due to a single attribute (e.g., gender), but to the combination of some of them (e.g., race, gender, poverty), which requires a multi-attribute solution (Ghai et al., 2021; Roy et al., 2021; Williams, 2014). The need for an awareness of SDOH is crucial to this kind of solution.

Do cultural biases exist in AI systems?

The same can also be said of the cultural biases that may be contained in algorithms developed by AI systems. A cultural bias in an AI system may occur when its design and/or use

produce a strong inclination either in favor or against some cultural group due to, among other factors, their cultural beliefs, customs, values, or religion. These factors can be considered as SDOH and, on occasion, as the cause of unfair inequalities in health (Chaturvedi et al., 2011).

As an example of the above, we know that diabetes is associated with a greater probability of aggravating the situation of a person infected with COVID-19 (Ortega et al., 2021), so an algorithm that is used in triage with the aim of maximizing the survival of COVID-19 patients will probably incorporate diabetes among the risk factors to make a prognosis of survival in the ICU. Although other factors such as health insurance, medication costs, and physician-related attitudes play an important role when designing a diabetes treatment plan, insulin therapy is a fundamental part of diabetes treatment. However, there are cultural elements that cannot be reduced to socio-economic factors (as they are cultural assumptions about health that are present in the whole cultural group, regardless of their economic status) and that strongly influence the use of insulin therapy. Some of these cultural factors act as barriers and contribute to an underutilization of insulin. Thus, for example, in the United States, some African Americans perceive that insulin causes organ damage (Aikens and Piette, 2009), while the belief that insulin causes macrovascular and microvascular complications such as blindness, damage to the kidneys or pancreas, or even death is also common among Hispanics and other minority groups (Aikens and Piette, 2009). The same belief occurs among Asian patients in Singapore (Wong et al., 2011). On the other hand, fasting is part of Muslims' religious beliefs, and, in the Canadian context, this sometimes affects their decisions about insulin therapy due to potential interference with their religious obligations (Visram, 2013).

Beyond diabetes, obesity or a propensity for vascular diseases, to give just a few examples, all of which are risk factors for patients worsening after infection by COVID-19, there may be cultural circumstances beyond the control of individuals that condition their treatment and evolution. In such cases, if an AI system incorporates only biological indices to assess the health and survival of a COVID-19 patient, it can seriously harm people and social groups whose cultural beliefs prevent them from having better control of their health. The conception of health also often responds to normative criteria and not only biological ones (Nordenfelt, 2006; Venkatapuram, 2011), which leads to a critical *ex ante* review of what conceptual basis algorithmic systems are based on.

The definition of health used by the system designers may have an impact on different parts of the lifecycle for developing an ML algorithm (Casacuberta et al., 2022). Therefore, it makes sense to think that algorithms producing a strong bias against people who hold certain cultural beliefs are producing biases that can lead to algorithmic discrimination and injustice. Further studies are needed to

better understand the nature of such biases and their presence in AI systems in order to mitigate them when they appear and, if possible, eliminate them.

Ethically good and bad biases

The fact that we generally consider biases as being negative suggests that the notion of *bias* is inherently normative. However, are all biases *morally* wrong? In other words, can some machine biases in Medical AI that favor particular individuals or groups be considered permissible or even desirable? This is another issue that has received scant attention. One standard approach is to differentiate bad biases from good biases by saying that the former lead to misdiagnosis or treatment errors, while the latter lead to a *justified* differential treatment. This is somewhat unspecific, however. On what sort of relevant grounds can a bias be considered justified? Without any further qualification, the distinction between good and bad biases remains difficult to make (Starke et al., 2021).

Mirjam Pot et al. (Pot and Prainsack, 2021; Pot et al., 2021) offer a more useful perspective. They criticize the view that ML biases are mainly systematic distortions and misrepresentations of the population, this implying that biases are a mere technical problem to be solved by technological means. In fact, biases are also a socio-political problem, in that they may underpin or undermine health inequities (i.e., unjust inequalities in health status between individuals or populations). In other words, training ML with more data or using better models will not always correct the underlying inequities (Pot et al., 2021). Pot *et al.* therefore propose evaluating the valence of biases according to their impact on social injustices in health. Thus, an ML bias is “bad” if it increases health inequities, and “good” if it reduces them. They go on to argue that creating deliberate biases may be desirable as long as they have beneficial equity effects, such as including historically marginalized groups (Pot et al., 2021). Although this view might be somewhat controversial, it opens the door for discussing the ethics of algorithmic affirmative action—something we believe should be further explored in the future. More generally, it introduces an important distinction related to the normativity of biases. On the one hand, their epistemic normativity, which relates to the following question: Does the bias contribute to better describing and predicting the world? And on the other, their moral and political normativity: Do biases contribute to making the world a better place? Biases in medical AI are not only scientifically relevant due to their being a distortion of reality, but they are also morally and politically relevant because of their impact on health inequities.

Measures to prevent the appearance of bias in AI

In order to prevent and mitigate the appearance of biases and other ethical problems such as those presented in the

previous section, organizations and institutions have deemed it necessary to implement regulatory frameworks and ethical guidelines to help AI developers and designers create a more ethical AI.

Governance and regulatory framework

The desire to mitigate risks in AI has led different transnational organizations and institutions to establish frameworks, guidelines and recommendations aimed at generating a sustainable ecosystem for technological development. With the idea of promoting a safe and reliable development of AI, the European Union has created different groups of experts, including the Ad hoc Committee on AI (CAHAI), appointed by the European Commission, and the group of AI experts belonging to the Organization for Economic Cooperation and Development (OECD). These organizations aim to ensure the technology's compliance and alignment with human rights, respect of dignity and certain ethical values such as transparency, security, and privacy, among others.

More than 75 organizations—including governments, companies, academic institutions, and NGOs such as Amnesty International—have produced documents with high-level guidelines in this respect (Jobin et al., 2019). Besides the formation of the groups of experts mentioned above, the European Commission has also begun to publish different regulatory frameworks that are helping to shape the development of AI in Europe. The White Paper on AI (European Commission, 2020) has the goal of promoting the uptake of AI and addressing risks associated with certain uses of the technology. This document also includes a draft regulation to ban some systems that are deemed unacceptable, such as biometrical surveillance in public space or systems classified as high risk due to their inherent biases.

Further, in April 2021, the European Commission published a framework called Regulation of the European Parliament and of the Council: Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (*EUR-Lex—52021PC0206—EN—EUR-Lex*, no date) to try to unify the governing of AI technology in the EU in addressing ethical and human rights concerns. The aims of this regulation are to guarantee that systems are safe and respect the regulation and values of the EU, improve governance, guarantee legal certainty and facilitate the development of a single market.

More than 70 recommendations and guidelines have been formulated on ethics and AI in recent years (Floridi, 2019), covering issues such as agency, autonomy, morality, rights, trust, transparency, and so on. Although this shows intense work and considerable interest in the ethical dimension of AI, unfortunately it also generates a duplication of work, confusion, and noise (Robles Carrillo, 2020). Thus,

the ethics of AI may become a mere “whitewash” if mechanisms are not established for compliance and implementation of the ethical recommendations proposed. Worst of all is the use of ethics by large technology corporations as an “alibi” to avoid regulation (Ochigame, 2019).

However, it is not just a matter of “correcting” the biases of certain databases, but of looking beyond that and considering the structural issues, historical antecedents, and power asymmetries involved in algorithmic injustice (Birhane, 2021). In the specific case of medical algorithms, we would be talking about the “social determinants of health”: “Algorithmic systems never emerge in a social, historical, and political vacuum, and to divorce them from the contingent background in which they are embedded is erroneous” (Birhane, 2021: 8). While there is no mandatory regulation, the scope of these recommendations is short and does not address the balance of power between those who develop AI systems and the communities subject to it, especially those that have been racialized historically.

Recently, private companies have started to spring up dedicated to the ethical analysis of algorithms for the detection, prevention and mitigation of bias and discrimination. However, the challenge remains to develop independent public bodies that monitor the development and implementation of algorithms, especially in very sensitive areas related to fundamental rights (e.g. the management of social rights, benefits, or penal attributions). This is the context for the Spanish government's creation of the Spanish Artificial Intelligence Supervisory Agency (AESIA). This agency will audit algorithms used by social networks, public administrations and companies with the aim of “minimizing significant risks to people's health and safety, as well as to their fundamental rights, that may arise from the use of artificial intelligence systems.” It will be important to analyze how this agency evolves, whether it will be able to capture the plurality and complexity of this field, and whether it will be able to escape market pressures. In any case, it is a worthy project for AI governance.

Assessment tools for the mitigation of bias in AI systems

Mitigating bias in AI systems has become one of the most important goals for a just and equitable development, deployment and use of these systems. The uncontrolled use of AI systems in the prevention, mitigation and monitoring of the COVID-19 pandemic has highlighted the importance of finding tools to help mitigate the appearance of biases that can lead to other ethical concerns and technical problems.

Algorithmic auditing characteristically assesses the consistency or robustness of the technological system design or the outcomes of the system. Two kinds of audits are

especially relevant to our purpose here: *functionality* audits, which focus on the rationale behind the decision, for example code audits entail reviewing the source code; and *impact* audits, which investigate the effects of an algorithm's outputs (Mökander and Floridi, 2021). Ethical-based auditing has gained in prominence recently, since it helps to identify, visualize and communicate the embedded values in a system and allows stakeholders to identify who should be accountable for potential ethical damages (Mökander et al., 2021).

Besides the categorization of functionality versus impact audits, another distinction can be made that is relevant for this discussion. Ethical assessment can have either a retrospective or a prospective focus regarding the impact and effects of a system, and too often, ethical reflection in audits is the former, that is involving assessments that are conducted ex post to diagnose and resolve a problem that already exists as such (perhaps prior to implementation). In other words, the aim is to find biases in *existing* systems to prevent these systems from causing harm. While detecting problems before they affect others is both necessary and beneficial, some have contended that this

approach does not take full advantage of the true value of a commitment to ethical reflection and action during the whole design process, and not only at its end. Contrarily, a prospective focus, even using principles and similar tools such as checklists, seeks to prevent biases from occurring by focusing on the decision-making process during the design stages, and therefore not only on correcting issues after the system has already been designed. To exemplify this, Beard and Longstaff (2018) proposed a framework for the design of technology that goes beyond offering standards to pass the “sniff test,” but rather seeks to assist and inform designers and developers to ensure that their design avoids harm and contributes to the good right from the very early conceptual stages of a design process. In sum, if retrospective assessment seeks to fix a system that is biased, prospective assessment is forward-looking and seeks to prevent biases from creeping in at all. Naturally, both approaches can coexist and be beneficial in bias mitigation.

Based on our analysis, we can offer a first list of approaches aimed at guiding and assessing the design and development of AI systems (Table 2). We should

Table 2. Tools for the mitigation of bias in AI systems.

Ethical checklists	<p><i>Goal:</i> To provide guidance through an evaluation that uses different questions related to the ethical aspects of the design, development, and deployment of an AI system. Evaluation and guidance checklists tend to cover all the milestones in an AI project and generally cover ethical values such as fairness, privacy, safety and robustness, security, and other social impacts.</p> <p><i>Description:</i> In general, ethical checklists consist of binary questions (Yes/No answers), which can oversimplify very complex decisions with various competing factors in deceptively simple compliance processes (Madaio et al., 2020). The format of ethical checklists is influenced by checklists used in software development, which can lead to the wrong idea that ethical problems can sometimes be resolved by purely technical solutions.</p> <p><i>Examples:</i> Microsoft's AI Fairness Checklist focuses mainly on the values of equity and discrimination throughout the design and implementation process (Leslie, 2019). The Turing Institute has also compiled an assessment checklist in Understanding Artificial Intelligence Ethics and Safety (Leslie, 2019), which is integrated within its own responsible AI framework. The High-Level Expert Group on Artificial Intelligence has also developed its own Assessment List for Trustworthy Artificial Intelligence (ALTAI), a practical tool that helps organizations, institutions and companies evaluate the reliability of its AI systems (ALTAI—The Assessment List on Trustworthy Artificial Intelligence Futurium, no date).</p>
Ethical standards	<p><i>Goal:</i> To ensure the ethical soundness of a process and its results.</p> <p><i>Description:</i> Standards are documents that formalize expertise around a subject by providing formal recommendations on how to do something in order to adhere to that standard (i.e., making a product or managing a process).</p> <p><i>Examples:</i> Standards such as those proposed by the Institute of Electrical and Electronics Engineers (IEEE) (Olszewska, 2020) or the International Standard Association (ISO).</p>
Ethical certifications	<p><i>Goal:</i> To qualify a certain AI system through a quality seal that helps users identify the risks of using the technology.</p> <p><i>Description:</i> The evaluation might be based on codified standards such as those referred to above, or it might be based on heuristic or qualitative evaluations by experts based on ad hoc criteria.</p> <p><i>Examples:</i> The Institute of Electrical and Electronics Engineers (IEEE) has implemented its own ethical certification, called IEEE CertifAIEd, which verifies the system's compliance with a series of ethical criteria (transparency, accountability, appearance of bias and privacy) and whose objective is to inspire confidence in the AI system. However, the most widespread certification is the VCIO Model (Values, Criteria, Indicators, Observables) (<i>AI Ethics Impact Group: From Principles to Practice—VDE</i>, no date) established by the AI Ethics Impact Group (AIEI Group). This model proposes a universal evaluation system for AI applications based on an ethical scoring system. The Group proposes six key values that serve as a benchmark for evaluation and comparison; they are transparency, responsibility, privacy, justice, reliability, and environmental sustainability. Each value is composed of</p>

(continue)

Table 2. continue

Ethical toolkits	<p>several measurable indicators that define a rate. This system also includes a risk matrix to evaluate the degree of dependency and potential damage an AI can cause.</p> <p><i>Goal:</i> To allow developers and designers to carry out ethical assessments of their systems either during design and development, or once the system is designed but before shipping and deployment.</p> <p><i>Description:</i> These toolkits are usually based on broad practical guides that allow self-evaluation of the system.</p> <p><i>Example:</i> An example of assessment that focuses on the shipping and deployment stages is Aequitas, which is an open source bias auditing tool developed by the University of Chicago Center for Data Science and Public Policy. The system helps audit the prediction of risk and produces a report with the outcome (Saleiro et al., 2018).</p>
Ethical guidelines	<p><i>Goal:</i> To address ethical concerns during the design stages.</p> <p><i>Description:</i> Ethical guidelines are developed by groups and organizations to define ethical standards for a responsible use of AI. They provide insights and recommendations that function as key references for designers and developers.</p> <p><i>Example:</i> An example of ethical guidelines is the Ethics guidelines for trustworthy AI developed by the High-Level Expert Group, which includes a pilot version of an AI Assessment list to operationalize a reliable AI. This assessment list aims to provide a basic process for self-evaluation. It is divided into different questions, summarized in the following categories: (1) Human agency and oversight; (2) Technical robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination and fairness; (6) Societal and environmental well-being; and (7) Accountability</p>

emphasize that this categorization is tentative, however, and might also overlap at times (for instance, a toolkit might include checklists or a certification might be based on an evaluation that uses standards).

Conclusions

To sum up, the existence of bias in AI systems used for triage and risk prediction implemented during the COVID-19 pandemic can result in discrimination and algorithmic injustice (Delgado et al., 2022). Nevertheless, different strategies and legal frameworks are being developed, both for the design and in the use of this type of system, to prevent and avoid the appearance of bias. That said, even with these resources, there is a risk that these strategies will be incomplete if they do not incorporate the crucial perspective of the SDOH. SDOH can be a source of bias and must be considered along with other biases such as race, age, gender or disability.

Therefore, given the proliferation of biases in AI systems, the developers of such systems and health care policy-makers should include a plurality of profiles above and beyond purely technical ones, such as experts from other fields from the social sciences and humanities (e.g. anthropology, sociology, ethics, or gender studies). The inclusion of experts from these fields could increase the epistemic diversity needed to rethink and tackle AI biases during the COVID-19 pandemic and beyond. Finally, from the perspective of algorithmic governance, it would also be desirable to establish an independent entity capable of reviewing and analyzing these systems from multiple perspectives.

Acknowledgements

The authors would like to thank Joaquín Hortal for his insights, and Barnaby Griffiths for the revision of the manuscript and the two anonymous reviewers for their thoughtful comments.

Authors' contributions

A.d.M. and J.D. contributed to the design of the manuscript. I.P.J. took over the revisions after peer-review. All co-authors have identified the main aspects of the study. All co-authors have discussed all the relevant aspects of the paper. J.D. and A.d.M. have created the tables and figures. All the co-authors critically reviewed the article, and contributed to the writing and editing process, as well as the review. All the co-authors have approved the final manuscript.

Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been funded thanks to the "Ayudas Fundación BBVA a Equipos de Investigación Científica SARS-CoV-2 y COVID-19" in Humanities.

ORCID iDs

Alicia de Manuel  <https://orcid.org/0000-0003-3195-9970>

Janet Delgado  <https://orcid.org/0000-0002-3681-8571>

Txetxu Ausín  <https://orcid.org/0000-0003-2098-5540>

References

- Abrams EM and Szeffler SJ (2020) COVID-19 and the impact of social determinants of health. *The Lancet. Respiratory Medicine* 8(7): 659.
- Afifi RA, et al. (2020) "Most at risk" for COVID19? The imperative to expand the definition from biological to social factors for equity. *Preventive Medicine* 139: 1–4.
- AI Ethics Impact Group: From Principles to Practice—VDE (no date) Available at: <https://www.ai-ethics-impact.org/en> (Accessed: 30 March 2022).

- Aikens JE and Piette JD (2009) Diabetic patients' medication underuse, illness outcomes, and beliefs about antihyperglycemic and antihypertensive treatments. *Diabetes Care* 32(1): 19–24.
- ALTAI—The Assessment List on Trustworthy Artificial Intelligence | Futurium (no date) Available at: <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence> (Accessed: 5 February 2022).
- Amann J, et al. (2020) Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20(1): 1–9.
- Beard M and Longstaff S (2018) Ethical by Design: Principles for Good Technology.
- Birhane A (2021) Algorithmic injustice: A relational ethics approach. *Patterns (New York, N.Y.)* 2(2): 1–9.
- Casacuberta D, Guersenzvaig A and Moyano C (2022) Justificatory explanations in machine learning: For increased transparency through documenting how key concepts drive and underpin design and engineering decisions. *AI & Society*: 1–15.
- Chaturvedi S, et al. (2011) Are we reluctant to talk about cultural determinants? *The Indian Journal of Medical Research* 133(4): 361. Available at: <http://pmc/articles/PMC3103166/> (Accessed: 5 February 2022).
- Choraś M, et al. (2020) Machine learning—The results are not the only thing that matters! what about security, explainability and fairness? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: 615–628. doi: 10.1007/978-3-030-50423-6_46.
- Chu Ch., et al. (2022) Ageism in artificial intelligence: A review. *Innovation in Aging* 6(1): 663.
- Cossette-Lefebvre H and Maclure J (2022) AI's fairness problem: Understanding wrongful discrimination in the context of automated decision-making. *AI Ethics*: 1–15. doi: 10.1007/s43681-022-00233-w.
- Council of Europe (no date) Ai and control of COVID-19 Coronavirus. Available at: <https://www.coe.int/en/web/artificial-intelligence/ai-and-control-of-covid-19-coronavirus#> (Accessed: 08 February 2023).
- Danks D and London AJ (2017) Algorithmic bias in autonomous systems. *International Joint Conference on Artificial Intelligence* 17: 4691–4697.
- Delgado J, et al. (2022) Bias in algorithms of AI systems developed for COVID-19: A scoping review. *Journal of Bioethical Inquiry* 19(3): 407–419.
- EUR-Lex—52021PC0206—EN—EUR-Lex (no date). Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206> (Accessed: 30 March 2022).
- European Commission (2020) On artificial intelligence—A European approach to excellence and trust White Paper on Artificial Intelligence A European approach to excellence and trust. Available at: https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf. (Accessed: 30 March 2022).
- Floridi L (2019) Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy and Technology* 32(2): 185–193.
- Frykberg ER (2005) Triage: Principles and practice. *Scandinavian Journal of Surgery* 94(4): 272–278.
- Gao Y, et al. (2020) Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications* 11(1): 1–10.
- Ghai B, Hoque MN and Mueller K (2021) Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–7.
- Goggin G and Ellis K (2020) Disability, communication, and life itself in the COVID-19 pandemic. *Health Sociology Review* 29(2): 168–176.
- Grote T and Keeling G (2022) Enabling fairness in healthcare through machine learning. *Ethics and Information Technology* 24(3): 1–13.
- Healthy People 2020 (no date). Available at: <https://www.healthypeople.gov/2020/> (Accessed: 31 March 2022).
- Hedden B (2021) On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs* 49(2): 209–231.
- Hortal-Carmona J, et al. (2021) La eficiencia no basta. Análisis ético y recomendaciones para la distribución de recursos escasos en situación de pandemia. *Gaceta Sanitaria* 35(6): 525–533.
- Iserson KV and Moskop JC (2007) Triage in medicine, part I: Concept, history, and types. *Annals of Emergency Medicine* 49(3): 275–281.
- Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.
- Korinek A and Stiglitz JE (2021) COVID-19 driven advances in automation and artificial intelligence risk exacerbating economic inequality. *BMJ* 372: n367.
- Kostick-Quenet KM, et al. (2022) Mitigating racial bias in machine learning. *Journal of Law, Medicine & Ethics* 50(1): 92–100.
- Leslie D (2019) *Understanding Artificial Intelligence Ethics and Safety. A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. doi:10.5281/zenodo.3240529.
- Leslie D, et al. (2021) Does “AI” stand for augmenting inequality in the era of COVID-19 healthcare? *BMJ* 372: 1–5.
- Livingston M (2020) Preventing racial bias in federal AI. *JSPG* 16(2): 1–7.
- Luengo-Oroz M, et al. (2021) From artificial intelligence bias to inequality in the time of COVID-19. *IEEE Technology and Society Magazine* 40(1): 71–79.
- Madaio MA, et al. (2020) Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. Conference on Human Factors in Computing Systems—Proceedings. doi: 10.1145/3313831.3376445.
- Marjanovic O, Cecez-Kecmanovic D and Vidgen R (2021) Theorising algorithmic justice. *European Journal of Information Systems* 31(1): 1–19.
- Marmot M, et al. (2012) WHO European review of social determinants of health and the health divide. *Lancet (London, England)* 380(9846): 1011–1029.
- Mezza I (1992) Triage: Setting priorities for health care. *Nursing Forum* 27(2): 15–19.
- Mittelstadt BD and Floridi L (2016) The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341.
- Mökander J and Floridi L (2021) Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31(2): 323–327.

- Mökander J, et al. (2021) Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics* 27(4): 1–30.
- Moseley D (2021) Bias. In: LaFollette H (ed.) *The International Encyclopedia of Ethics*. New York: John Wiley & Sons, Ltd, pp. 1–6. doi: 10.1002/9781444367072.WBIEE861.
- Naseem M, Akhund R, Arshad H, et al. (2020) Exploring the potential of artificial intelligence and machine learning to combat COVID-19 and existing opportunities for LMIC: A scoping review. *Journal of Primary Care & Community Health* 11: 2150132720963634.
- Noor P (2020) Can we trust AI not to further embed racial bias and prejudice? *BMJ* 368: 1–3.
- Nordenfelt L (2006) Establishing a middle-range position in the theory of health: A reply to my critics. *Medicine, Health Care and Philosophy* 10(1): 29–32.
- Noseworthy PA, et al. (2020) Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis. *Circulation: Arrhythmia and Electrophysiology* 13: 7988.
- Ochigame R (2019) The invention of ‘Ethical AI’. How Big Tech Manipulates Academia to Avoid Regulation.
- Olszewska JI (2020) IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being: IEEE Standard 7010-2020.
- Ortega E, et al. (2021) Risk factors for severe outcomes in people with diabetes hospitalised for COVID-19: A cross-sectional database study. *BMJ Open* 11(7): 1–10.
- Paremoer L, et al. (2021) COVID-19 pandemic and the social determinants of health. *BMJ* 372: 1–5.
- Pasquale F (2021) Promoting data for well-being while minimizing stigma. In: Moore M and Tambini D (eds) *Regulating Big Tech: Policy Responses to Digital Dominance* (New York, online edition, Oxford Academic). 180–C10.P74.
- Pfeiffer RN and Gail MH (2011) 基因的改变 NIH public access. *Biometrics* 67(3): 1057–1065.
- Pot M and Prainsack B (2021) Reply to Letter to the Editor on “Not all biases are bad: Equitable and inequitable biases in machine learning and radiology”. *Insights into Imaging* 12(1): 1–2.
- Pot M, Kieusseyan N and Prainsack B (2021) Not all biases are bad: Equitable and inequitable biases in machine learning and radiology. *Insights into Imaging* 12(1): 1–10.
- Quiroz-Juárez MA, et al. (2021) Identification of high-risk COVID-19 patients using machine learning. *PLOS ONE* 16(9): e0257234.
- Robles Carrillo M (2020) Artificial intelligence: From ethics to law. *Telecommunications Policy* 44(6): 101937.
- Röösli E, Rice B and Hernandez-Boussard T (2021) Bias at warp speed: How AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association* 28(1): 190–192.
- Roy A, Iosifidis V and Ntoutsi E (2021) Multi-fair pareto boosting. Available at: <https://arxiv.org/abs/2104.13312v2> (Accessed: 5 February 2022).
- Saleiro P, et al. (2018) Aequitas: A bias and fairness audit toolkit. Available at <https://arxiv.org/abs/1811.05577v2> (Accessed: 5 February 2022).
- Soler W, et al. (2010) Triage: A key tool in emergency care. *Anales del Sistema Sanitario de Navarra* 33(SUPP1): 55–68.
- Starke G, De Clercq E and Elger BS (2021) Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy* 24(3): 341–349.
- Steyerberg EW, et al. (2010) Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* 21(1): 128.
- Tat E, Bhatt DL and Rabbat MG (2020) Addressing bias: Artificial intelligence in cardiovascular medicine. *The Lancet Digital Health* 2(12): e635–e636.
- Turner Lee N (2018) Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* 16(3): 252–260.
- Venkatapuram S (2011) Health justice: An argument from the capabilities approach. *Polity* 88: 643–645.
- Visram H (2013) Patient barriers to insulin use in multi-ethnic populations. *Canadian Journal of Diabetes* 37(3): 202–204.
- Wilkinson RG and Marmot MG (2003) Social determinants of health: The solid facts. Available at: [https://books.google.co.jp/books?hl=es&lr=&id=QDFzqNZZHLMC&oi=fnd&pg=PA5&dq=Wilkinson,+R.+/+Marmot,+M+\(eds\).+Social+Determinants+of+Health.+The+solid+Facts.+Geneve:+World+Health+Organization,+2003.&ots=xWrIcIVRpw&sig=ODfVG1azItTRDnKbskwIb8HFkVA#v=onepage&](https://books.google.co.jp/books?hl=es&lr=&id=QDFzqNZZHLMC&oi=fnd&pg=PA5&dq=Wilkinson,+R.+/+Marmot,+M+(eds).+Social+Determinants+of+Health.+The+solid+Facts.+Geneve:+World+Health+Organization,+2003.&ots=xWrIcIVRpw&sig=ODfVG1azItTRDnKbskwIb8HFkVA#v=onepage&) (Accessed: 5 February 2022).
- Williams JC (2014) Double jeopardy? An empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law & Gender* 37: 185–242.
- Williams JC, et al. (2020) Colorblind algorithms: Racism in the era of COVID-19. *Journal of the National Medical Association* 112(5): 550–552.
- Wang L, et al. (2021) Artificial intelligence for COVID-19: A systematic review. *Frontiers in Medicine* 8: 704256.
- Wong S, et al. (2011) Perceptions of insulin therapy amongst Asian patients with diabetes in Singapore. *Diabetic Medicine: A Journal of the British Diabetic Association* 28(2): 206–211.