AI and Ethics (2022) 2:771–787
https://doi.org/10.1007/s43681-022-00138-8

ORIGINAL RESEARCH

AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry

Lorenzo Belenguer[1]

Received: 28 August 2021 / Accepted: 11 January 2022 / Published online: 10 February 2022

Abstract
A new and unorthodox approach to deal with discriminatory bias in Artificial Intelligence is needed. As it is explored in detail, the current literature is a dichotomy with studies originating from the contrasting fields of study of either philosophy and sociology or data science and programming. It is suggested that there is a need instead for an integration of both academic approaches, and needs to be machine-centric rather than human-centric applied with a deep understanding of societal and individual prejudices. This article is a novel approach developed into a framework of action: a bias impact assessment to raise awareness of bias and why, a clear set of methodologies as shown in a table comparing with the four stages of pharmaceutical trials, and a summary flowchart. Finally, this study concludes the need for a transnational independent body with enough power to guarantee the implementation of those solutions.

Keywords AI · Ethics · Algorithmic bias · Discrimination · Machine learning

1 Introduction

This essay explores the highly pertinent topic of bias within artificial intelligence (AI). Attempting to move understanding beyond the existing philosophical debates, this study bases itself within the emerging field of Applied Ethics. In recent years, researchers in this discipline have highlighted and created debate around potential issues surrounding AI such as regarding data privacy or discriminatory outcomes. They have also been instrumental in devising novel solutions to such dilemmas, creating ethical frameworks intended to enhance the rapidly evolving technology-based solutions present in every corner of modern life.

Existing literature analysing AI bias tend to originate from one of two, very separate, academic spheres. On one side, the theories are formed from a philosophical or sociological perspective, which study problems either existing or expected in the future. Whilst useful in creating debate, these tend to present either no solutions at all or overly simplified single solutions [13, 15, 19, 29, 32, 35, 56, 60, 64, 89, 96, 104]. On the other hand, it is the approach by data scientists and programmers that characterise AI biases as *bugs* implying that it is just a technical issue like security that needs to be fixed [Tramer et al. 2016, 53, 54]. We need a combination of both approaches within a clear framework of action (Fig. 1).

This essay seeks to identify whether an approach, combining these two dominant academic fields of study may create a more successful solution in reducing AI bias. How can abstract ideas, such as fairness or social justice, be translated into applicable ethical frameworks? Then, into coding understandable by a machine? This study will analyse the value of a set of tools, focussed on solving bias, adapted or inspired by the policies of the pharmaceutical companies.[1] Such industries have a long history of developing risk-assessment methodologies, on a stage-by-stage basis, facing the known and the unknown. The pharmaceutical industry also has a long history of Applied Ethics,[2] which will be explored. In addition, they have adopted an independent

✉ Lorenzo Belenguer
Lorenzo.Belenguer@gmail.com

[1] London, UK

[1] For a complementary uptake, please see [73] report.

[2] The pharmaceutical industry is far from perfect, but it is in a better position now than when eugenics experiments were openly conducted on underprivileged sectors of society with no consequences. Today there are mechanisms to take a pharmaceutical company to Court if harm to society is proven as the over-promotion of opioids derivatives in the US, for example. Such legal mechanisms are underdeveloped or non-existent in the AI industry.

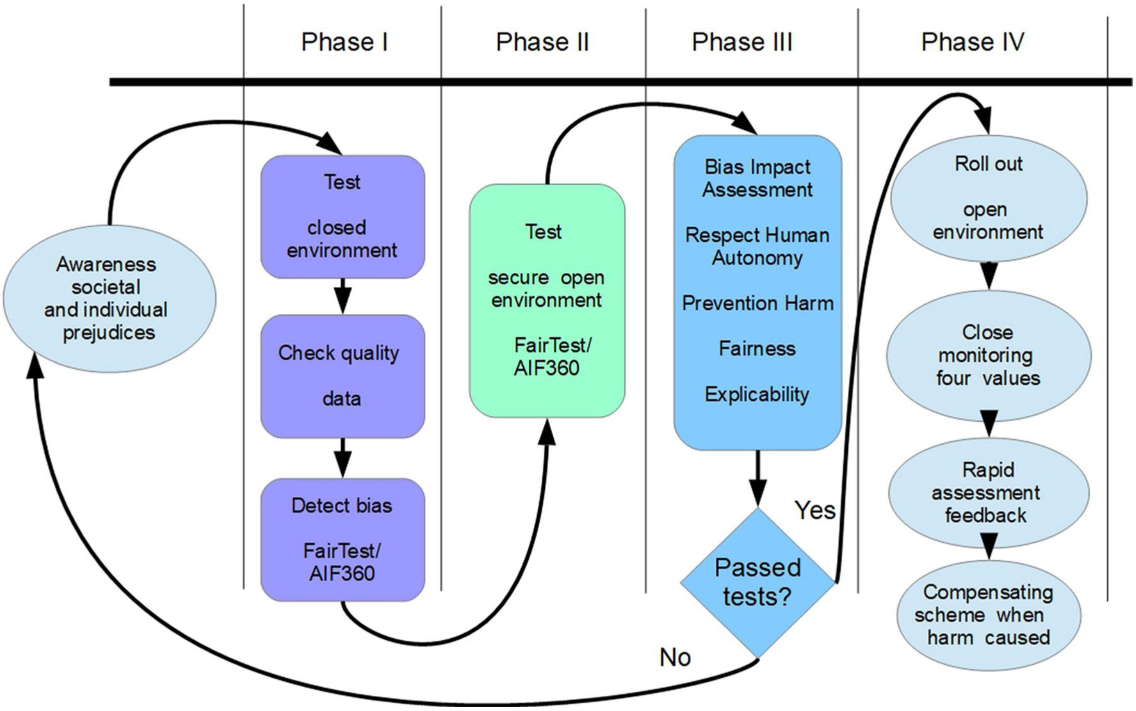

Phase I
Phase II
Phase III
Phase IV
Awareness societal and individual prejudices
Test closed environment
Check quality data
Detect bias FairTest/ AIF360
Test secure open environment FairTest/ AIF360
Bias Impact Assessment
Respect Human Autonomy
Prevention Harm
Fairness
Explicability
Passed tests?
Yes
No
Roll out open environment
Close monitoring four values
Rapid assessment feedback
Compensating scheme when harm caused

Fig. 1 This is a summary flowchart of a framework of action that I suggest in this article. All definitions and actions will be further explained in the next sections. Actions in phase I, phase II, and phase III can be conducted in a different order according to individual needs except the final test. As the technologies evolve, some actions might need to be expanded or added. AI bias framework of action (summary). Lorenzo Belenguer

regulatory body (US FDA, UK MHRA or EU MDA)—a necessity that keeps coming up in many AI Ethics discussions [36].

A case will be created highlighting the discrimination issue in algorithmic decision-making using two case studies which clearly show the presence of well-documented biases (based on race and gender) with the application of a suggested model to conduct a bias impact assessment. In the subsequent sections, the problems associated with data collection will be introduced, suggesting three possible tools (four-stage implementation, boxing method and a more practical application of the protected groups' concept). Finally, the study will explore the potential of an independent regulatory body with enough power to guarantee implementation and what this could mean for the future of AI.

Finally, to reiterate the need for machine-centric solutions, as Computer and Information Science professors Kearns and Roth [49, p. 21] note:

> "Of course, the first challenge in asking an algorithm to be fair or private is agreeing on what those words should mean in the first place—and not in the way a lawyer or philosopher might describe them, but in so precise a manner that they can be "explained" to a machine".

2 Definition of artificial intelligence, machine learning, algorithms and AI bias

Artificial Intelligence is a central theme of this study and as such it is first important to clarify what this means. Norvig and Russell, the authors of *Artificial Intelligence: A Modern Approach,* considered one of the seminal textbooks on AI, provide a comprehensive definition of AI [72, p. viii]:

> "The main unifying theme is the idea of an intelligent agent. We define AI as the study of agents that receive percepts from the environment and perform actions. [...] We explain the role of learning as extending the reach of the designer into unknown environments".

As this definition suggests, the main concept of AI is of an intelligent agent that develops the capacity of independent reasoning. To achieve that goal, and through specific actions, the non-human agent needs to collect information, find ways to process that data, and benefit from the act of learning to reach further than the role of its designer into *unknown environments.*

To achieve those results, some of the most successful approaches that machines use are Machine Learning models, or ML, which consist of training and data. ML is an attempt to mimic one of the ways humans learn. For example, if an adult wants to explain a sports car to a child, it can compare it with a standard car to develop an understanding on an already built system of knowledge by the child. A common example is to provide the machine with labelled photos of cats and dogs, and afterwards show unlabelled photos of both of these animals so the machine develops a system of reasoning to differentiate which is which. This is an example of supervised learning, which is one of the three main approaches explained below [3].

Machine-learning models can have the capacity to evolve, develop and adapt their production in accordance with training information streams [3]. The models can share their newly acquired expertise with other machines using techniques as part of what it is called model deployment. As Opeyemi [66] defines: "Model deployment […] refers to the arrangement and interactions of software components within a system to achieve a predefined goal".

Influenced by the categorisations proffered by Murphy [57] and Alpaydin [3], machine learning can be divided into three main approaches:

1. Supervised learning: when the data given to the model are labelled. For example, image identification between dogs and cats with the images labelled accordingly.
2. Unsupervised learning: when the machine is given raw unlabelled data and tries to find patterns or commonalities. An example could be data mining on the internet when the algorithm looks for trends or any other form of useful information.
3. Reinforcement learning: when the machine is set loose in an environment and only occasionally receives feedback on the outcomes in the form of punishment or reward. For example, in the case of a machine playing a game like chess.

Deep learning is a subset of ML that uses artificial neural networks (or ANNs) as the backbone of their model with a minimum of three layers of depth to process the information [40]. ANNs can be compared with how the brain cells form different associational networks to process information. ANNs can be very powerful as they have the capability to be flexible and find new routes in the neural networks to better process data—similar to the human brain (Fig. 2).

The word *algorithm* and its study come from a Persian mathematician from the ninth century called al-Khwarizmi (the term derives from his name) [58]. At its basis, an algorithm is a set of instructions or rules that will attempt to solve a problem.

AI Bias is when the output of a machine-learning model can lead to the discrimination against specific groups or individuals. These tend to be groups that have been historically discriminated against and marginalised based on gender, social class, sexual orientation or race, but not in all cases. This could be because of prejudiced assumptions in the process of developing the model, or non-representative, inaccurate or simply wrong training data. It is important to highlight that bias means a deviation from the standard and does not necessarily lead to discrimination [38, p. 1]. For example, it can show differences in statistical patterns in the data collected like the different average height between human adults in relation to gender.

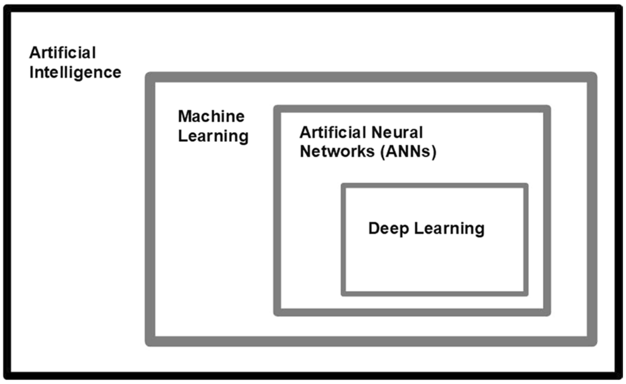


Fig. 2 This is a simplified diagram of where they fit in AI. Inspired by [40, p. 9]

Bias in data can show in many different ways which can lead to discrimination. This a non-comprehensive list that shows some of the most common type of bias that needs to be dealt with [54] and Suresh et al. [81]:

1. Historical bias. Historical bias is the type of bias that already exists in society and the collection of data reflects that.
2. Representation bias. Representation bias happens from how we define and sample from a population. For example, a lack geographical diversity in datasets like ImageNet (a large visual database designed for use in visual object recognition software research such as facial recognition) is an example for this type of bias [81]. This demonstrates a better representation of the pale skin population in the Western countries.
3. Measurement bias. Measurement bias happens from how we choose, analyse, and measure a particular feature. An example of this type of bias was demonstrated in the recidivism risk prediction tool COMPAS, which is one of the two cases studies evaluated in the article.
4. Evaluation bias. Evaluation bias happens during model evaluation. It includes the use of either disproportionate or inappropriate benchmarks for evaluation of applications. These benchmarks can be used in the evaluation of facial recognition systems that were biased towards skin colour and gender [23, 60].

5. Simpson's paradox. Simpson's paradox [14] can bias the analysis of heterogeneous data that consists of subgroups or individuals with different behavioural patterns. According to Simpson's paradox, a trend, association, or characteristic observed in underlying subgroups may be quite different from one subgroup to another.
6. Sampling bias. Sampling bias arises when the sampling of subgroups is non-randomised. It means that the trends estimated for one population may not be extrapolated to data collected from a new population.
7. Content production bias. Content Production bias occurs from structural, lexical, semantic, and syntactic differences in the contents generated by users according to age and gender groups among other characteristics [63].
8. Algorithmic bias. Algorithmic bias is when the bias is not actually in the input data and is created by the algorithm [71].

This article, as it is common in AI ethics literature, will concentrate on the problematic cases in which the outcome of bias may lead to discrimination by AI-based automated decision-making environments and an awareness of the different types can be helpful.

3 Algorithmic decision-making that discriminates and the problem with data

Algorithms rely on data, and their outcomes tend to be as good as the data provided and labelled and the way the mathematical formulations are devised. Even in an unsupervised ML model working with raw data, the machine might find discriminatory societal patterns and replicate them. The computer can be used as a proxy for a human, relinquishing them of any moral responsibility [60].

Humans can be biased as it is the way society is constructed and maintained by a minority elite at the top of the hierarchy. This elite constantly develops strategies, either consciously or unconsciously, to prevent others from accessing their privileges [17]—and the elaboration of prejudices is one of them.[3] As Noble [60, p. 14] explains in her influential book: "Part of the challenge of understanding algorithmic oppression is to understand that mathematical formulations to drive automated decisions are made by human beings". The machines incorporate those prejudices, becoming a proxy of humans and delegating responsibility.

The process of data mining,[4] one of the ways algorithms collect and analyse data [88], can already be discriminatory as a start, because it decides which data are of value, which is of no value and its weight—how valuable it is. The decision process tends to rely on previous data, its outcomes and the initial weight given by the programmer. One example can be when the word *woman* was penalised, by being given a negative or a lower weight, on a CV selection process based on the data of an industry traditionally dominated by men like the tech industry [79, 82]. The outcome ended discriminating women in the selection process [33].

Some ML models, like supervised learning, learn by examining previous cases and understanding how data are labelled, which is called training. Training data that are biased can lead to discriminatory ML models. It can happen in two ways [84]:

1. A set of prejudicial examples from which the model learns or in the case of under-represented groups which receives an incorrect or unfair valuation.
2. The training data are non-existent or incomplete.

While there are many reasons for incomplete or biased data, two are particularly relevant: historical human biases and incomplete or unrepresentative data [51]. Societies, as described by Bonilla-Silva [17], are structured by an elite at the high end of the hierarchy, controlling power and the top stages in the decision-making processes (e.g. judges, senior civil servants and politicians), whose biases have the ability to be adopted as a standard across society which then lead to historical human biases.[5] The second reason, incomplete or unrepresentative data, is a consequence of the first. Some data from specific groups' databases can be either non-existent or simply incorrect, as was initially the case with female car drivers who were a minority. When the first safety belts and airbags for cars were designed, they suited tall males (the average engineer). Any other humans with other physical characteristics, especially shorter stature, were not considered, ending in a higher fatality rate in a car crash [18].

The quality of the collected data will influence the quality of the algorithmic decisions. If the data are biased with one example being prejudicial bias—racial bias being a well-documented case [2, 23], the outcome will likely follow suit unless appropriate controls are put in place. There must be an evaluation of the quality and response accordingly before applying the algorithm. There is always a trade-off between accuracy, fairness and privacy that needs to be taken into

[3] Prejudices and abuse of power occur in all directions and among members of the same social class. However, I am more interested in elite discrimination from the top to the bottom of the social scale as it affects bigger sectors of the population and the monopoly of the implementation of discriminatory ML models on a larger scale.

[4] The ethical issues of Web Data Mining are well explored in this paper Van Wel et al. [88].

[5] Not that it is that simple or the only reason. However, it is an important factor.

account as Corbett-Davies et al. [31] examined in their study. For example, how much data, and private data, we need to gather to detect bias in cases when the data are sensitive, like cancer patients receiving the right health insurance product.

3.1 What a bias impact assessment is and how to develop one

The bias impact assessment can be very helpful in clearly identifying the main stakeholders, their interests and their position of power when blocking or allowing necessary changes and the short- and long-term impacts. The concept *fairness through awareness* was introduced by Dwork et al. [35], and it states that if we wish to mitigate or remove bias in an algorithmic model, the first step is to be aware of the biases and why they occur. The bias impact assessment does that and hence its relevance.

The assessment can also provide a better understanding of complex social and historical contexts as well as supporting a practical and ethical framework for spotting possible bias. It can then help with how to mitigate them in the assessment of automated decision AI systems and facilitates accountability. Qualitative and quantitative measures could be given for a range of categories determined by the ethical framework, which would include: bias, privacy, security, psychological well-being, among other ethical concerns. Reisman et al. [70, pp. 5–6] demonstrate the necessity of algorithmic impact assessment in AI as standard practice: "Impact assessments are nothing new. We have seen them implemented in scientific and policy domains as wide-ranging as environmental protection, human rights, data protection, and privacy […] scientific and policy experts have been developing on the topic of algorithmic accountability" (also see [25]).

For the two case studies, the bias impact assessment will be conducted within two frameworks: the analysis by the experienced scholar on AI Ethics Dr Sarah Spiekermann,[6] *Ethical IT innovation: A value-based system design approach* (2015), and the K7 conditions from the white paper on trustworthy AI published by the High-Level Expert Group on AI of the EU [45]. However, Spiekermann's model will be the primary focus, and the K7 conditions from the EU white paper will play a secondary role. The main reason is that Spiekermann's model follows clear steps in identifying key elements such as stakeholders, benefits and harms while being complemented by the four value approach from the High-Level Expert EU paper.

Summaries of the key steps taken are as follows:

The first step, called value discovery, consists of naming the stakeholders affected, how those benefits or harms map to values.

The second step called value conceptualisation is the process of breaking down harms and benefits into their constituent parts.

The third step, empirical value investigation, is when we differentiate the stakeholders' views and priorities.

The fourth and final step, the technical value investigation, is how to increase the benefits and minimise or eliminate harm.

However, the model will be simplified by reducing the first three steps into naming the stakeholders, their benefits, harms, priorities and interests. It will not develop into the fourth step as I will be presenting some solutions in the following sections. The key concepts will be further explained while carrying on the case studies as it is the easiest way to understand them. However, as the concept of values can be quite abstract, it is helpful to provide a list of four values, which facilitates a robust analysis to detect bias, from the EU white paper on Trustworthy AI, 2019 p. 14[7]:

1. Respect for human autonomy. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills.
2. Prevention of harm. AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity.
3. Fairness. The development, deployment and use of AI systems must be fair. The substantive dimension implies a commitment to ensuring equal and just distribution of benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
4. Explicability. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected.

3.2 Algorithmic decision-making that discriminates based on race

Correctional Offender Management Profiling for Alternative Sanction, known as COMPAS, is a predictive ML

[6] Dr Spiekermann is a co-chair of IEEE's first standardisation effort on ethical engineering (IEEE P7000). She has been published in leading IS and CS Journals including the Journal of Information Technology, the IEEE Transactions on Software Engineering, Communications of the ACM, and the European Journal of IS, where she served as Editor until 2013 (obtained from IEEE, Institute of Electrical and Electronics Engineers, website).

[7] As this article focuses on bias AI, I will prioritise the values that affect bias.

Fig. 3 (Human 1/black male) left, prior offence: 1 resisting arrest without violence, given a high risk assessment of 10. Subsequent offences: none. (Human 2/white male) right, prior offence: 1 attempted burglary, given a low risk assessment of 3. Subsequent offences: 3 drug possessions. COMPAS. Source Angwin et al. [2]

model designed to provide US courts with defendants recidivism risks scores that oscillate between 0 and 10. It predicts how likely the defendant is to re-offend by perpetrating a violent crime, taking into account up to 137 variables [61] such as gender and age and criminal history, with a specific weight given to each. COMPAS is a risk-assessment tool that aids the operations of criminal justice organisations and is an extension of other judicial information systems [2]. For example, an individual who scores 8 has twice the re-offending rate of those who have 4. Defendants waiting for a trial with a high-risk score are more likely to be imprisoned while waiting for trial than those with low risks, so the consequences of a wrong assessment can be dire. Someone can be wrongly imprisoned while awaiting trial who would not re-offend while a more dangerous individual more likely to offend would be let free (Fig. 3).

Northpointe, renamed Equivant, the company that created COMPAS, claimed that they do not use race as one of the factors. However, a study of defendants in Broward County, Florida, showed that black individuals are much more likely to be classified as high risk [2]. The same paper indicates that black people who did not re-offend were twice as likely to be classified as high risk compared to a white person as the risk score assessment in Fig. 4 indicates.

The first step in the model for a bias impact assessment is called value discovery:

1. Judges, police officers and other members of staff in the Justice and Police department—they benefit by imprisoning individuals who are likely to re-offend and freeing individuals who are not likely to re-offend to keep costs down and effectively invest resources.
2. Defendants—they would expect a fair trial, being treated with dignity, access to a competent lawyer and assistance with rebuilding their lives.
3. Prison institutions in the US—private prison facilities, including non-secure community corrections centres and home confinement, held 15% of the federal prison population on December 31, 2017 [21, p. 16]. Their business model operates on the basis of more prisoners, more profit [24]. The private sector has an incentive to encourage incarcerating as many people from lower class backgrounds with restricted access to lawyers who are less likely to legally challenge unfair treatment. The public sector, operating state prisons, seems to be willing to maintain the status quo by the figures provided in point 5.
4. Society as a whole—it needs to feel safe by keeping serious offenders in prison while facilitating re-integration of non-violent offenders.
5. Minorities, especially from the Black community, seem to be the victims of racial injustice. According to the World Prison Brief [100], the US has one of the highest incarceration rates in the world. In 2018, the figure was 23.4% of the total population. Black adults make up 33% of US prison population while just making 12% of the US adult population [102].

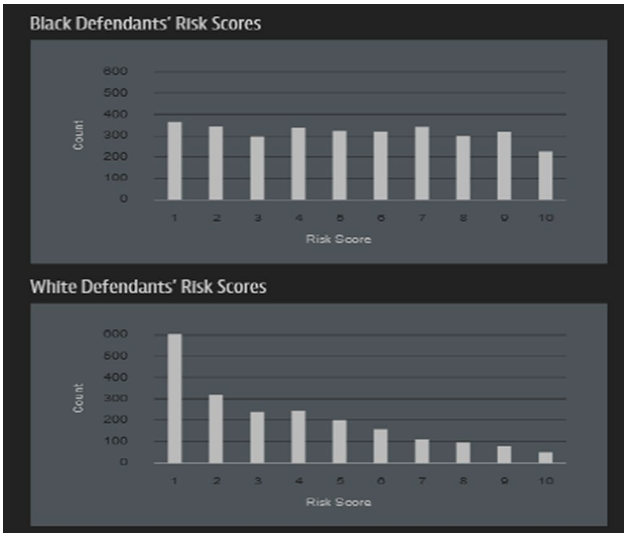


Fig. 4 These charts show that scores for white defendants tended toward lower-risk categories. Scores for black defendants did not. Source: ProPublica analysis of data from Broward County, Florida. Angwin et al. [2]

The second step is called value conceptualisation:

There seems to be an imbalance of power between a privileged white population that holds a majority of high-ranking positions in the Justice and Police departments, which could favour institutionalised racism over the black population,[8] as figures seem to demonstrate in the case of the black population being over-represented in the prison population [22]. The elite have the benefit of reinforcing privileges, and the

[8] To simplify and more data available, I have not mentioned the Latinx community and other communities that also endure discrimination based on race.

rest of the population have limited access to progress to well-paid jobs and colleges [8]. There is a tension between fairness, one of the key values, and the tendency to maintain the status quo, which might be based on generational held prejudices against other groups, according to Bell [8]. It seems to contradict two of the main aims of applying Justice: prevention of harm and respect for human autonomy. Incarceration needs to be executed as a last resort. Finally, this model does not fully explain how it calculated those risk scores, so explicability seems non-existent [2].

The third step is empirical value investigation:

I have made an initial distinction between professionals in the Justice, Prisons and Police institutions and individuals who commit offences at various levels. However, according to the statistics, it is evident that it is a socio-economic issue in which race plays a big part. Prisons and police enforcement seems to be a tool to perpetuate classism and maintain a rigid social structure, of which racism is a by-product [17, 22].

The ML model COMPAS collects historical data from previous discriminatory court sentences and enhances those prejudices, with the added characteristic of being a proxy of a human and delegating moral responsibility.

3.3 Algorithmic decision-making that discriminates based on gender

In 2014, Amazon started to use an algorithm to select the top five CVs from one hundred applicants. The model ended in penalising the word *woman* and favouring male applicants. Although it was removed as soon as the bias was detected, and the company states that it was never in use, it is a good illustration of gender discriminatory outcomes as the case study will demonstrate [33, 91].

It is a problematic finding because Amazon has a long history in the use of algorithmic decision-making, as users have long been recommended products based on previous searches and purchases [52]. AI has been at the heart of their business for years and they are hence assumed to be at the forefront of such technologies.

Initially, it was a great idea to receive a selection of the top five applicants saving time and energy in a process that could be automated. The algorithm applied the patterns in selecting individuals from the last 10 years, and it simply replicated that. Indeed, tech companies have one of the largest gender disparities in all industries. Female programmers in IT constitute only 19.8% of the total workforce [79, p. 5], and make up only a quarter of employees in the technology industry [82], p. 1]. Following those patterns, Amazon's system learns how to reinforce those normalised discriminatory outcomes. One strategy was very straightforward by penalising the word woman. Any CV that contained this word or any others denoting the female gender of the applicant, like attending a women's college, for example, downgraded the score, according to people familiar with the project [33, 91].

There were some attempts to ameliorate the problem, but the issue of gender discrimination was deep-rooted. The term *woman* and other words with similar connotations were not taken into account to facilitate neutrality. However, there was no guarantee of other discriminatory issues not coming out. The system was already based on a rather biased database. It proved challenging to remove bias and guarantee equal opportunities, so Amazon decided to scrap it altogether [33].

Although Amazon said it was never implemented, it did not confirm that recruiters had no access to the machine's recommendations. Thus consciously, or unconsciously, affecting the selection process.

The first step in the model for a bias impact assessment is called value discovery:

1. CEO and top managerial positions—it is in their benefit to recruit the best people and not to be reported as a gender-biased company. After all, around 50% of the population are women, and it is not a good idea to upset such a significant percentage of the market.
2. HR department—although they are expected to recruit the best candidate, a tool that can do your job easier is tempting. If rather than scanning 100 CVs, the officer only needs to go through five, this is an attractive option despite not garnering the most desirable results.
3. The rest of the staff—if the team is predominantly male, some members might wish to keep it like that. There is a tendency in a male-dominated industry for some members of the staff to be apprehensive of a more gender-balanced working group [5, 75]. This may lead to HR being encouraged to continue in one direction that suits them.
4. The candidates—a well-suited candidate would feel dispirited by not having an interview opportunity just because of belonging to the wrong gender. Other candidates might appreciate it, although they might be unaware of the process being discriminatory. Many candidates would not like to work for a company with such discriminatory practices.

The second step is called value conceptualisation:

In this specific example, the disadvantaged group are the women,[9] as they are blocked or impeded from accessing those jobs and limiting respect for human autonomy (financial independence). In addition, there is no

[9] Many other groups might have been treated unfairly, such as Latino or black males, but I will concentrate on gender discrimination in this case study.

prevention of harm (self-esteem/self-value), if women apply for those positions and it stops their career development. Thirdly, a lack of fairness, if a candidate matching those requisites is not selected for a job interview because of their being the wrong gender. Finally, there is no explicability of why two candidates with the same characteristics, except gender, are not invited to the interview. There is a tension between a team that needs to display the diversity of talent in a contemporary society and the tendency to maintain the status quo in an industry historically dominated by men. A less diverse source of ideas and backgrounds might result in poorer creativity and overall innovation in producing new products. A diverse team benefits the company in the sense that all voices are represented and their needs tailored [59].

The third step is empirical value investigation:

The tech industry is male dominated, comprising almost 75% of the workforce [79, 82]. Some of the male workers would prefer business as usual.[10] They might be prejudiced against women and prefer the perpetuation of those values, making some women feel unwelcome in tech companies. Wajcman [92] argues in her book of the perceived *masculinity* of technology. Other male workers, and the rest of the female workers, would prefer a more gender-balanced company where everybody feels welcome and the management board reflects that gender diversity [59, 101].

Both case studies conclude with the necessity of applying a bias impact assessment to raise awareness of any bias and its possible reasons before being implemented as demonstrated by Dwork et al. [35]—and its urgency. The New York University professor Onuoha [65] uses the term *algorithmic violence* in her concerns to capture the "ways an algorithm or automated decision-making system inflicts [violence] by preventing people from meeting their basic needs", such as a fair trial or job selection.

4 Possible machine-centric solutions adapted from or inspired by the pharmaceutical industry

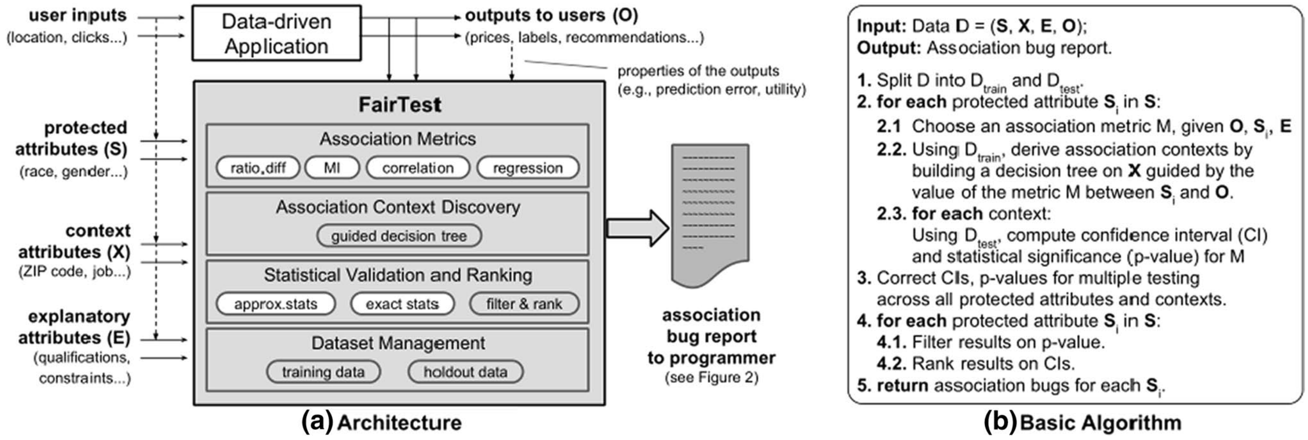
The pharmaceutical industry has a long history in applied Ethics and risk-assessment methodologies in a multidisciplinary field, which is advantageous in finding ethical solutions.[11] It also has a variety of collecting and contrasting data strategies like randomised control trials, which can help improve bias impact assessments by unearthing unexpected outcomes. In addition, they conduct their trials on different age, gender and other characteristics groups at different stages and compare results. This process helps to develop a standard methodology to maximise benefits and remove, or minimise, harm. A further result is achieving effective methods to measure those outcomes. Examples of a standardised set of core outcome measures can be the development of the design of machines focussed on measuring patients' health status by analysing blood tests. In the case of AI, as is demonstrated later, it can be ML models created just for the purpose of measuring (bias in those examples) such as FairTest (Fig. 5) or AI Fairness 360. All those methodologies are regulated by an independent body such as the US FDA or UK MHRA. The central aim is to understand what can be learned from pharmaceutical companies which can be applied to machine-learning models. The AI industry is similar to the pharmaceutical industry in its multidisciplinary environment, its need for diverse voices and expertise, and the pivotal role that applied Ethics has played in its development. As Santoro [74, p. 1] explains: "Perhaps no business engages the worlds of science, medicine, economics, health, human rights, government, and social welfare as much as the pharmaceutical industry".

Pharmaceutical companies were not expected to conduct trials to demonstrate the safety and accuracy of their medical products until 1962. That year, the US Congress passed the Kefauver-Harris Amendments to the Food, Drug and Cosmetics act of 1938, and Europe followed suit soon afterwards [74, p. 12]. The AI industry seems to enjoy a similar path of unregulated growth followed, hopefully, by regulations, better awareness in the industry and by the mainstream market—although it is vying to achieve this in a shorter period. That does not mean that there are no public safety issues like the anti-inflammatory drug Vioxx produced by Merck, which was linked to heart attacks and strokes in long-term use and withdrawn from the market in 2004 [10, 74, p. 13]. However, side effects are usually noticed and recorded with the assistance of surveillance and monitoring systems like Pharmacovigilance in phase IV (Hauben et al. 2009), and there is a regulatory procedure to act upon it. Unfortunately, these safety measures do not seem to exist in the AI industry.

The four possible solutions adapted from the pharmaceutical industry that I am going to discuss in this article are boxing methods (as adapted from the four stages implementation), blind testing (inspired by testing on different groups), a better application of the protected groups' concept (as vulnerable groups), and a regulatory body (such as the US FDA or UK HMRA) where I will try to make a case for one at a transnational level. This combination of approaches and methodologies can result in a robust analysis and implementation of solutions in one applicable set (Table 1).

[10] Whitehouse et al. [97] draws on survey data to examine horizontal and vertical gender segregation within IT employment in Australia. Not all data can be extrapolated to other countries and cultures, and it may be outdated. However, tech culture is global and it is an example of blocking women in IT jobs due to the *masculinity* of technology [92].

[11] Pharmaceutical companies' business model is based on profit, but there are regulatory procedures to minimise harm, remove products when proven harmful and compensate the victims which do not exist in the AI industry.



```
Input: Data D = (S, X, E, O);
Output: Association bug report.
1. Split D into $D_{train}$ and $D_{test}$.
2. for each protected attribute $S_i$ in S:
    2.1 Choose an association metric M, given O, $S_i$, E
    2.2. Using $D_{train}$, derive association contexts by
         building a decision tree on X guided by the
         value of the metric M between $S_i$ and O.
    2.3. for each context:
         Using $D_{test}$, compute confidence interval (CI)
         and statistical significance (p-value) for M
3. Correct CIs, p-values for multiple testing
   across all protected attributes and contexts.
4. for each protected attribute $S_i$ in S:
    4.1. Filter results on p-value.
    4.2. Rank results on CIs.
5. return association bugs for each $S_i$.
```

Fig. 5 FairTest architecture. **a** Grey boxes denote FairTest components. Rounded boxes denote specific mechanisms. White rounded boxes denote extensibility points; transparent rounded boxes denote core, generic mechanisms. **b** FairTest's basic algorithm, realising the UA framework methodology. S, X, E denote protected, context, and explanatory attributes, respectively; O denotes the output quantity of interest [86, p. 10] (The reader, like me, is not expected to fully understand the complexities of an algorithmic model. The main reason is to have an overview of the process and its different steps.)

Table 1 A comparison of the four stages between both industries

	Pharmaceutical industry	AI industry
Phase I	Determining the safety of the product	Testing in a closed virtual environment, checking the quality of data, detecting bias (*FairTest/AIF360*)
Phase II	Testing its effectiveness	Testing in a secure open environment (*FairTest/AIF360*)
Phase III	Comparing its effectiveness with the standard treatment available	Bias impact assessment (checking four values: respect for human autonomy, prevention of harm, fairness, and explicability). In addition, protected groups and other individuals who might be affected
Phase IV	Monitoring any risks and benefits once the product is in the market	Ready to be used an open environment, close monitoring four values, feedback from affected groups, compensating scheme when harm caused to individuals

When we discuss the importance of regulating AI and the new technologies, another recurrent argument is that it might delay innovation, become costly and would harm consumers. The same argument was used when new chemicals were invented. They ended up harming the environment and the surrounding population, at a great cost cleaning up the mess and compensating the victims, which could have been avoided by implementing safety measures. For example, DuPont, a highly respected company that produced the popular material Teflon used in cooking tools, caused environmental damage that ended up costing the company around a billion dollars [78, p. 1]. Not doing the right thing can end up harming a business. The welfare of humans, other living creatures, and the environment needs to be prioritised over any possible unchecked innovation. On the other hand, regulation does not have to affect innovation, for example, when pharmaceutical companies developed a COVID-19 vaccine utilising a cutting-edge technology, mRNA, in a record time (Kim 2020).

4.1 A more effective way to apply the protected groups' concept

As with the concept of vulnerable groups in the pharmaceutical industry testing process, we all have an idea of what might constitute a protected group. Protected Groups are defined by the Equality Act 2010 as: "a group of persons defined by reference to a particular characteristic against which it is illegal to discriminate". There are nine protected characteristics identified: age, disability, sex, marriage and civil partnerships, pregnancy and maternity, race, religion and belief, sexual orientation and gender reassignment (The National Archives).

Moreover, there was a hope that ML models would remove those prejudices as machines executed the process, but the two case studies demonstrate that the opposite is true. The data used in training is intrinsically biased, as society is, and it tends to simply replicate human behaviour—one of the core aims of Artificial Intelligence.

A more effective use of this concept is tested by Wang et al. [94] in their paper when introducing the idea of noisy protected groups. By "noisy protected groups", the authors mean when data are corrupted, missing or unreliable due to social pressures. The participants may be withholding or providing false information to avoid retribution. For example, in a conservative society, a gay person might claim to be heterosexual to avoid homophobic attacks, so the data collected is unreliable. In those cases, the protected groups' data are unreliable, leading to an unreliable outcome. As they identify issues in the abstract, p. 1: "Many existing fairness criteria for machine learning involve equalizing some metric across protected groups such as race or gender. However, practitioners trying to audit or enforce such group-based criteria can easily face the problem of noisy or biased protected group information".

4.2 Boxing methods

Today, clinical trials are the norm. The drugs are tested on humans only after they have undergone laboratory testing. This takes the form of a series of successive clinical trials known as phase I, phase II, phase III, and phase IV trials. The access to the drugs is limited, boxed in, and opened up as it progresses through the different stages until being fully available in the market. Each phase of a clinical trial has a distinct objective. Phase I trials are conducted to determine the safety of the product, phase II trials test whether the drug is effective, phase III trials will compare its effectiveness with the standard treatment available, and phase IV trials monitor any risks and benefits once the product is in the market. As Sedgwick [77, p. 1] assures: "Drugs under development that are found to be unsafe or ineffective will not progress through all four phases". No medicine would be allowed to reach the market without a risk and safety report and seal of approval from the regulatory body and monitoring systems for a follow-up.

The benefit of AI is that most trials can be conducted in a virtual setting protecting the population from harm [53]. There does not appear to be any reasons why tech companies cannot conduct their businesses in a similar manner to mitigate bias. This standard procedure seems effective and easy-to-conduct to minimise harm, and better knowledge is acquired on possible side effects and effective doses in a clear, standardised manner (stages I, II, III and IV).

This process, which is standard procedure in the pharmaceutical companies, can be a great model to follow before rolling out any model that uses algorithmic decision-making techniques. The first benefit is the necessity of applied Ethics in the AI industry moral sphere, which still seems to be devoid of a moral compass. This process should be obligatory by legislation to incorporate ethics at the heart of designing any ML model as is commonplace in the pharmaceutical industry. Second, it provides an ethical framework with a clear set of instructions for programmers and data scientists to follow. Third, and finally, it facilitates a better understanding of the known effects and expected and unexpected outcomes that would be unearthed with the testing of the product. Some of the solutions have already been explained in the previous chapters and the rest will be developed in the final ones.

This is a simplified model focussed on bias which needs to be extended to other issues such as data privacy. Some of the concepts in this table will be explained in the next sections.

It makes sense to start Phase I from a virtual environment, where the algorithm can interact in total freedom. It is a good first step to assess any bias, malfunctions or adverse effects while access is ring-fenced. It is similar to the sandbox concept of testing in programming but much more complex as there are more variables in real life with the assistance of *FairTest* which will be explained later. For the same reason AI technologies are constantly improving, the same can be said of virtual environments used for testing purposes (see McDuff [53] as an example). An industry can be developed to provide those services.

As with any digital environment at this early stage, it needs to be inaccessible to external agents. The computer needs to be connected by cable to servers. In addition, the machine kept to a bare minimum of programming to avoid any pollution from unnecessary programming, bugs and possible undetected malware. However, it would need to be exposed at a later stage to what would be a typical environment in the outside world. Initially, it needs to be isolated and blocked from accessing outside information. Bostrom [19, p. 131] and Chalmers [29] develops a similar strategy, although not intentionally to mitigate bias, it can apply to achieve this goal too. As ML systems evolve into more sophisticated and autonomous agents, this initial testing would need to become a compulsory and more thorough process. We already have so much data available in digital format that it should be reasonably easy to simulate simplified, but effective, replicas of the world virtually and any variation or alternative space to conduct the initial test. In addition, it is imperative to avoid data that has not been tested thoroughly (as is explained in Sects. 3 and 4.3), to avoid any unknown bias. Part of the testing process would be to differentiate the essential dataset to train the ML model and filter out the less relevant or unnecessary data. It would limit abuse in the extraction of data. Once a standard set of safety measures is passed, then it is ready for the next step.

In the second stage, the ML model would need to interact with the real world, albeit in a limited capacity. It would consist of interactions with humans still in a physically or digitally isolated or limited environment. If it takes place in a digitally enclosed environment, a limited number of devices can be connected either wirelessly or if it is a very sensitive project only by cable. Trained testers will try to find bias using different methods like blind testing, which will be further explained later, or the way data are used in the training model. It is the stage when historically biased data can be detected and addressed accordingly. A report can be produced on inaccuracies, inconsistencies and other indicators of bias in the system [53]. It is perhaps the most critical stage as the next one is a rollout to the general population. For example, in 2015, it was found that Google image misidentified black persons as gorillas [41] or the two cases explained before where discriminatory outcomes were affected by race and gender, which should have been identified in the early stages prior to being launched to the market.

The bias impact assessment will be conducted in the third stage; as per the details in the previous section. Special relevance needs to be given to check the four values: respect for human autonomy, prevention of harm, fairness, and explicability. Finally, a clear identification of all the stakeholders, their interests and the tensions when those interests are not being met. It is the stage that an awareness of bias and why should be identified.

Once the concerns of the second and third testing stage are addressed, the system is ready for a general application on the whole or part, of the population—the fourth stage. The report will indicate bias in the system that might need to be attentively watched. A straightforward feedback tool needs to be in place to swiftly solve bias when detected by its users. Hopefully, those discriminatory bias can be detected at an earlier stage.[12] Finally, a compensating scheme, when harm caused to individuals, needs to be included to encourage compliance as it is the case in the pharmaceutical companies (Fleming [39] makes a good case for compensation plans in his paper).

4.3 Blind testing

Pharmaceutical companies test their products on different groups based on gender, ethnicity and age, amongst other factors, to unearth unexpected side effects that might not affect other groups of the population [77]. The main intention, at this stage, is to detect different outcomes from different groups and determine whether there is a fair reasoning behind this or whether the outcome is discriminatory, as seen in the COMPAS case study.

If we assess a mortgage application, all individuals with the same salaries, identical credit records, and other factors needed to evaluate the application should receive the same rate. Gender or race should not be a reason for being downgraded or upgraded (Bartlett 2019). When the outcome differs, a clear and unprejudiced reason needs to be made available.

This process facilitates identifying where the bias occurs, whether gender or race-related, or for any other reason. Once identified, it is much easier to correct the bias, perhaps by treating it as a protected group and giving it a particular weight. One example could be the COMPAS ML model used as a case study. This process of blind testing would unearth a disparity in outcomes related to race. Once this problem has been detected, the protected group, the black male population, can be identified and start adjusting the weights, full awareness of the disparities and a sensitive approach to reduce bias.

Many published papers have sought to examine and tackle this issue by testing the algorithm. For example, [86], provide the FairTest, a tool designed to test the ML model, assisting developers in checking data-driven applications to detect unfairness in outcomes. It is designed to investigate associations between application outcomes (such as prices or premiums) and sensitive user attributes (such as race or gender). Once detected, it provides debugging capabilities that help programmers solve the detected unfair effects. Tramer et al. [86, pp. 1 and 2] describe their Test themselves and its use as: "We report on the use of FairTest to investigate and in some cases address disparate impact, offensive labelling, and uneven rates of algorithmic error in four data-driven applications".

Tech companies that implement algorithms need to be accountable for any form of unfair treatment and act upon any discrimination as soon as it is spotted as it is often discussed in the AI Ethics fields [32, 60, 104]. This is the same treatment as when pharmaceutical companies discover a harmful side effect on any of their products (Fleming [39], Phase IV). There are clear guidelines for what to do next, such as removing the product from the market until proven safe for human consumption.

A concept of great interest to be introduced at this stage is Unwarranted Associations, UA, as it can be helpful to identify unfair associations when labelling data which might lead to bias. It is included in their FairTest toolkit under the UA framework [86]. In the same paper, p. 6, they define an unwarranted association as: "Any statistically significant association, in a semantically meaningful user subpopulation, between a protected attribute and an algorithmic output, where the association has no accompanying explanatory factor". *Explanatory factors* are the reasons that contribute to the outcome and could explain differences in results. For example, an algorithmic model that produces the patterns to make safety vests and one protected group, women, are found to receive an average smaller size. An explanatory factor would be that women tend to be physically smaller

[12] Although there are many other factors that need to be checked, like data privacy. In this article, I concentrate on bias. The main reason is to be able to introduce possible applicable solutions in a deeper manner.

than men. In this case, a group receives a different outcome, and the reason can be fully explained and is considered fair. On the other hand, this opens up the possibility for ill-use. For example, Google images were tagging images of black people with offensive, racist and incorrect remarks, on some occasions as gorillas [23]. As there is no satisfactory explanatory factor, then it is an example of an unwarranted association.

The first stage in the blind testing process offers additional safeguarding techniques: the UA framework, the association-guided tree construction algorithm, the design, implementation, and a thorough evaluation with FairTest.

The UA framework is the primary tool to discover and analyse bias associated with the data used to train the algorithm. As Tramer et al. [86] state on page 3: "Multiple primitives and metrics with broad applicability, explanatory factors, and fine-grained testing over user subpopulations, with rigorous statistical assessments, even for multiple adaptively chosen experiments". The first step has three key factors: testing, discovery, and error profiling identification which are part of the UA framework. A thorough examination of the labels to detect inconsistencies like when misidentifying a black person with another specie—as was the case in Google images is required. The association-guided tree construction algorithm further investigates the findings in the first step. It introduces a visual presentation to facilitate the interpretation and identify which subsets are affected by algorithmic bias [86]. It is instrumental to quickly identify the bias and inconsistencies found in the UA framework. The design, implementation, and evaluation of FairTest is the stage when those findings are translated into valuable coding for the machine. The source code will be provided as indicated in the study on page 3. Another source code available is AI Fairness 360 (AIF360), provided in the study [9] released under an Apache v2.0 license. This is a comprehensive package that includes a set of fairness metrics for datasets and models, including its explanations and algorithms to reduce bias. As they detailed in the paper: the initial AIF360 Python package implements inputs from 8 published papers, over 71 bias detection metrics, and 9 bias mitigation algorithms. It also includes an interactive Web experience.

This methodology can be very effective, especially in error profiling. For example, when Tramer et al. [86] tested *FairTest* to *Healthcare Prediction*, a winning approach from the Heritage Health Competition, they found out that there were errors when profiling some members of the elderly population, especially with some pre-existing health conditions [86, p. 4]. The report gave clear instructions on the steps to remove bias. Elderly people with some pre-existing health conditions were discriminated against incorrectly, predicting more visits to the hospital, in some cases, and being charged a higher premium. In a healthcare system that invests less money in black patients than white patients, the ML model feeding on that data can conclude that black patients are healthier than equally sick white patients and reduce the budget accordingly [62]. This is another example of a discriminatory outcome where tools like FairTest can be of great help.

Does the ML model facilitate an equal society? Are the conditions of protected groups improved over the years of applying the model? All ML models affect those issues in one way or another, and the current business models do not address these concerns with no way to legally enforce them. The three main possible solutions explained in this and the previous sections need an independent body with enough power to implement those measures. A bias-free AI is not achievable without an institution with enough power to guarantee compliance with those guidelines [26, 37, 70].

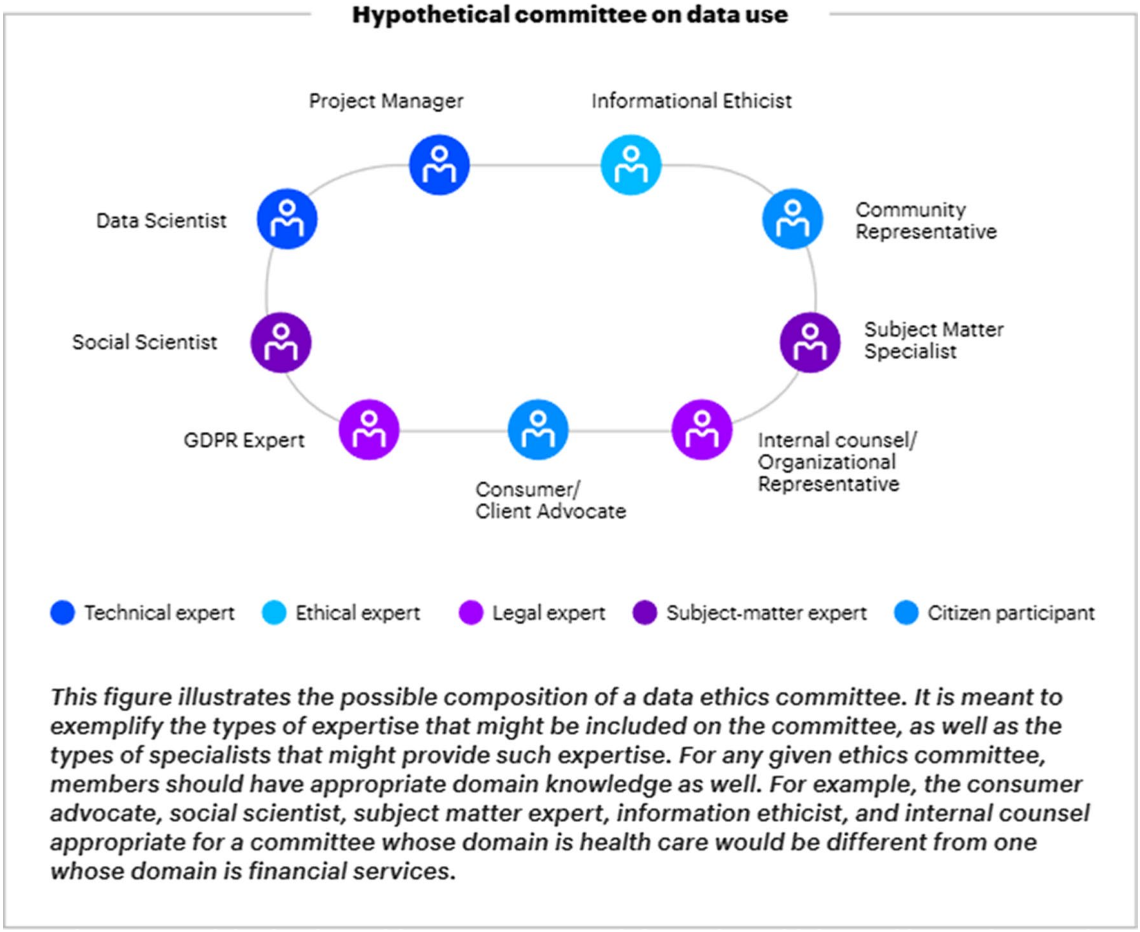
4.4 An independent regulatory body as transnational as possible

An independent body, on a transnational level, is needed. Or at least some kind of international coordination including as many countries as possible, is desirable. However, by its device-based nature, AI is a transnational technology, and it needs solutions applicable beyond borders. If we have a Nagasaki and Hiroshima nuclear incident in the AI industry, it will not be limited to a specific geographical area. The majority of the global population has smartphones, and to a limited extent, computers, the damage could be extended to the whole planet. In addition, tech companies like Google or Facebook operate beyond borders and its implications when things go wrong are transnational [99].

The body needs to be as multidisciplinary as possible, drawing from a diversity of expertise and backgrounds. Stakeholders need to represent all sectors of society. In the expertise area, it needs to cover: data science, applied ethics, coding, digital law and human rights activism. In the backgrounds area, it must be varied, especially on gender, race, sexual orientation and socio-economic class. Members of protected groups need to have a prominent role in guaranteeing fairness.[13] Those two initial demands should be part of a standard mindset of values and expertise [73].

AI has become so ubiquitous globally and powerful in a non-transparent way that setting up an independent body is an urgent necessity. AI can save lives, for example, in the case of cancer detection technologies in medical imaging [12]. It can improve our quality of life as it has done so in the past. An example can be when Gmail introduced an algorithm to remove spam emails [11]—the same principle could be used to reduce fake news. The potential benefits are vast as they could be its setbacks.

[13] Some may say that they need to have a more prominent role rather than just equal.



This figure illustrates the possible composition of a data ethics committee. It is meant to exemplify the types of expertise that might be included on the committee, as well as the types of specialists that might provide such expertise. For any given ethics committee, members should have appropriate domain knowledge as well. For example, the consumer advocate, social scientist, subject matter expert, information ethicist, and internal counsel appropriate for a committee whose domain is health care would be different from one whose domain is financial services.

Fig. 6 Sandler and Basl [73], p. 15

Its nature can be disruptive and profoundly affect an individual's future. For example, in the case of an undeserved incarceration while waiting for trial. Although the judge has the final say, algorithmic recommendations affect an overworked official's decision. The majority of the population does not have the expertise or awareness in relation to these types of harms. A regulatory body comprised of a diverse panel of experts can solve that problem (Fig. 6). Every time a new medicine is launched into the market, an independent body like the US FDA or the UK HMRA makes sure it is safe, or as safe as possible, and follows a clear set of guidelines [39, 43, 74, 77].[14]

The AI industry can strike a balance between innovation and regulation. There is no point in allowing an ML model to be deployed in an open environment if it harms people. Woolley et al. [99] provides case studies as examples such as: political bot intervention during pivotal events in Brazil or the origins of digital misinformation in Russia in which manipulation took place. As Reed [69, p. 2], Professor of Electronic Commerce Law at Queen Mary University of London, adds: "Good regulation would improve our perception of safety, and also our perception that humans remain in control. It could also mitigate any new risks that AI creates". Finally, good regulation mitigates harm, and it is more cost-effective than trying to cover the costs of damages and fines, as the previous example of DuPont shows.

An error in an ML model can easily affect several countries in the case of Amazon or Facebook. That is one of the main reasons for an international body, or at least, basic international guidelines. It could be called International Artificial Intelligence Organization (IAIO) as suggested by Erdelyi and Goldsmith [37, p. 5], as an intergovernmental organisation, and as they say: "to serve as an international forum for discussion and engage in standard setting activities".

[14] There are cases like the Boeing 737 MAX being in the market with faulty software and causing two fatal accidents. But that was caused by the lack of adequate monitoring of Boeing by the FAA, not by ineffective or inexistent regulation [44]. Commercial scheduled air travel remains among the safest modes of transportation (US National Safety Council 2019). Not perfect, but much better than unregulated.

A safety net that guarantees fundamental human rights is paramount. From that standard level, other countries might follow suit by imposing stricter regulations. It is the same case when a pharmaceutical company wishes to launch a new product and seeks approval per country or association of countries. Those independent bodies have developed similar guidelines. For example, a medicine approved by the FDA is very likely to be approved by other national bodies. The well-being of individuals can be improved on a global scale. Third world countries poor in resources can rely on the seal of approval from rich countries to allow a medicine to reach their market. In the general spirit of, 'if it is good enough for you, it is good enough for me'.

Legally binding regulation needs to give the independent body enough power to follow up on their recommendations to lawmakers to act upon them, effective implementation, transparency in the process, and tools to enforce those rules. This needs political will and citizens awareness[15] as well as a genuine intention by tech companies to change their business model. The independent body needs to be able to prosecute a breach once those recommendations become law and impose fines based on a percentage of the company's turnover. The fine needs to be high enough to act as a deterrent. Moreover, the percentage needs to be calculated on the turnover, rather than profit or the amount of tax paid, as international companies devise complex accounting mechanisms to avoid paying much tax (as becoming more common in the EU and US, Bageri 2013). The combination of strategies explained above can be a robust set of tools to guarantee the development of a trustworthy AI that benefits all members of society.

The argument for a more advisory role, or soft law, has been practised for years with hardly any progress. The time has come for a legally binding framework, albeit with mechanisms for flexibility and fast response to unexpected outcomes—either harmful or beneficial. Recommendations by advisory bodies tend to resort to vague language to accommodate all parties' interests. It tends to lead to nothing, as Hagendorff [42, p. 108] asserts: "In their AI Now 2017 Report, Kate Crawford and her team state that ethics and forms of soft governance "face real challenges" [26, p. 5]. This is mainly due to the fact that ethics has no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation between ethicists and individuals working in research and industry".

Finally, the benefits of transnational institutions can be demonstrated in the case of the European Union. Many laws and regulations have accelerated the fight for corruption, fiscal stability, accountability, Human Rights and fairness in members countries less willing to do so [47].

5 Conclusion

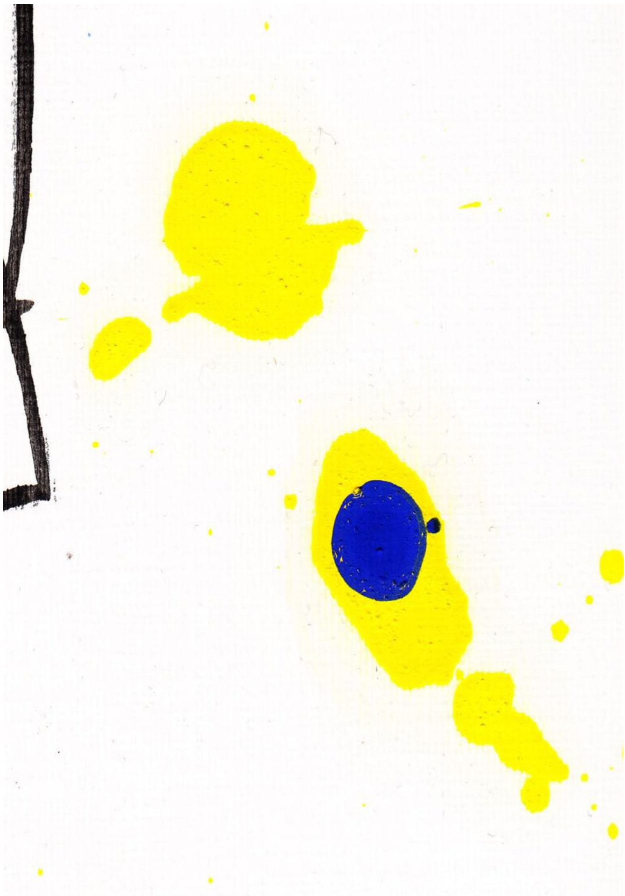
This article, in addition to analysing two discriminatory cases, has presented some possible solutions following, adapting, or being inspired by another industry with a long history of applying Ethics in its methodology to increase benefits and reduce, or remove, harm. As has been demonstrated, the pharmaceutical industries are far from perfect, but there are already expectations from consumers and governments which are not fulfilled yet by the AI industry and legally binding regulations when those expectations are not met. All these possible solutions are present, but are not collected as a framework of action as this article intends, and there is no guarantee by an independent body with the power to enforce them.

Studies by [8, 55, 67], are potent illustrators of embedded bias in society and the difficulty of removing them which is reflected in the AI industry. As Crawford [32, pp. 117–118] warns us: "The reproduction of harmful ideas is particularly dangerous now that AI has moved from being an experimental discipline used only in laboratories to being tested at scale on millions of people". We have now the technologies and awareness to at least mitigate, aiming to remove, bias. When scholars look back at discovering previous threats to humankind like climate change due to man-made pollution, as early as the 80s [90], they realise that providing the evidence is insufficient to modifying behaviour by companies and governments and they surmise that more pro-active strategies are needed. Data ethics need to be the core principle in developing any model in AI if we want fairness in a society of free citizens all enjoying equal fundamental rights in an egalitarian economic system as Hoffmann [46, p. 1] argues.

There are many challenges ahead if we want AI to be fair and bias-free (for a more detailed list see [54]. First, the concept of fairness and bias can mean different things for different people, it lacks uniformity although some basic principles, or values as previously described, can be agreed. Second, when the resources are shared, are we being equal? Equal in a sense that everybody is given the same level of resources, attention or receives the same outcome? If we concentrate on equity, are we distributing different amounts according to individual or group needs to achieve the same goal (protected groups such as people with disabilities). Equity and equality to mitigate bias might show conflicting results. Third, instances of unfairness in one group might not translate into another group, as it is in the previously explained Simpson's Paradox. For example, long waiting lists for a cancer treatment is considered unfair in the general population, but does not affect those with a private health insurance. Finally, the technologies in AI

[15] It is the reason why I have been advocating about the benefits of Citizens' Assemblies on AI to keep members of the Society informed and engaged. It could give politicians the public mandate to act upon it. Tech companies control the flow of information in the digital sphere with sophisticated algorithms. It is reasonable to suspect that they might interfere with accessing information that questions the technological status quo.

evolve so rapidly that new challenges and opportunities arise and additional methodologies might be needed such as with the adoption of quantum computing or 6G technologies. The time for action is now.



Still Life #4 (close up), oils on cotton paper, Lorenzo Belenguer.

Acknowledgements With immense gratitude to the editorial team for their great assistance, and the anonymous reviewers for their input. Secondly, to Ioannis Votsis, my MA dissertation supervisor, a truly vocational professor who provided me with superb insights and feedback. Thirdly, to Justine Seager for her great assistance in the initial editing. Finally, this paper is dedicated to the inspirational women and nonbinary of colour, especially Timnit Gebru and Joy Buolamwini, for pioneering a more diverse and inclusive approach to AI and Ethics.

References

1. Anderson, J., Rainie, L., Luchsinger, A.: Artificial intelligence and the future of humans. Pew Res. Center **10**, 12 (2018)
2. Angwin, J., et al.: (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016). Accessed 28 Mar 2021
3. Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
4. Bageri, V., Katsoulacos, Y., Spagnolo, G.: The distortive effects of antitrust fines based on revenue. Econ. J. **123**(572), F545–F557 (2013)
5. Bagilhole, B.: Being different is a very difficult row to hoe: survival strategies of women academics. In: Davies, S., Lubelska, C., Quinn, J. (eds.) Changing the Subject, pp. 15–28. Taylor & Francis, London (2017)
6. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
7. Bartlett, R., Morse, A., Stanton, R., Wallace, N.: Consumer-lending discrimination in the FinTech era. J. Financ. Econ. **143**(1), 30–56 (2022)
8. Bell, D. Faces at the Bottom of the Well: The Permanence of Racism. Hachette, UK (2018)
9. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S. AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943 (2018)
10. Bhattacharya, S. (2005). Up to 140,000 heart attacks linked to Vioxx. *New scientist*, *25*.
11. Bhuiyan, H., Ashiquzzaman, A., Juthi, T.I., Biswas, S., Ara, J.: A survey of existing e-mail spam filtering methods considering machine learning techniques. Glob. J. Comput. Sci. Technol. **18**(2-c)(2018)
12. Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., Mak, R.H.: Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin **69**(2), 127–157 (2019)
13. Binns, R.: Fairness in machine learning: lessons from political philosophy. In Conference on Fairness, Accountability and Transparency, pp. 149–159. PMLR (2018)
14. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)
15. Boddington, P.: Towards a Code of Ethics for Artificial Intelligence, pp. 27–37. Springer, Cham (2017)
16. Boden, M.A.: Creativity and artificial intelligence: a contradiction in terms. In: Paul, E., Kaufman, S. (eds.) The Philosophy of Creativity: New Essays, pp. 224–46. Oxford University Press, Oxford (2014)
17. Bonilla-Silva, E.: White Supremacy and Racism in the Post-Civil Rights Era. Lynne Rienner Publishers, Boulder (2001)
18. Bose, D., Segui-Gomez, S.C.D.M., Crandall, J.R.: Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk. Am. J. Public Health **101**(12), 2368–2373 (2011)
19. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. Camb. Handb. Artif. Intell. **1**, 316–334 (2014)
20. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
21. Bronson, J., Carson, E.A.: Prisoners in 2017. Age **500**, 400 (2019)
22. Brewer, R.M., Heitzeg, N.A.: The racialization of crime and punishment: criminal justice, color-blind racism, and the political economy of the prison industrial complex. Am. Behav. Sci. **51**(5), 625–644 (2008)
23. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp. 77–91. PMLR (2018)
24. Burkhardt, B.C.: Who is in private prisons? Demographic profiles of prisoners and workers in American private prisons. Int. J. Law Crime Just. **51**, 24–33 (2017)
25. Calvo, R.A., Peters, D., Cave, S.: Advancing impact assessment for intelligent systems. Nat. Mach. Intell. **2**(2), 89–91 (2020)

26. Campolo, A., Sanfilippo, M., Whittaker, M., Crawford, K.: AI now 2017 report. https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf (2017). Accessed 7 May 2021
27. Carrie, J.: More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru. *The Guardian.* https://amp.theguardian.com/technology/2020/dec/04/timnit-gebru-google-ai-fired-diversity-ethics (2020). Accessed 4 May 2021
28. Castelvecchi, D.: Can we open the black box of AI? Nat. News **538**(7623), 20 (2016)
29. Chalmers, D.: The singularity: a philosophical analysis. In: Schneider, S. (ed.) Science Fiction and Philosophy: From Time Travel to Superintelligence, pp. 171–224. Wiley, UK (2009)
30. Collingridge, D.: The Social Control of Technology. Frances Pinter (Publishers), London (1982)
31. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806 (2017)
32. Crawford, K.: The Atlas of AI. Yale University Press (2021)
33. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (2018). Accessed 24 Apr 2021
34. Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Meunier-FitzHugh, K.L., Meunier-FitzHugh, L.C.L., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., Williams, M.D.: Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int. J. Inf. Manag. **57**, 101994 (2019)
35. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
36. Erdélyi, O.J., Goldsmith, J.: Regulating Artificial Intelligence: Proposal for a Global Solution. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)
37. Erdélyi, O. J., Goldsmith, J.: Regulating artificial intelligence proposal for a global solution. Preprint at arXiv:2005.11072 (2020)
38. Ferrer, X., van Nuenen, T., Such, J.M., Coté, M., Criado, N.: Bias and discrimination in AI: a cross-disciplinary perspective. IEEE Technol. Soc. Mag. **40**(2), 72–80 (2021)
39. Fleming, J.G.: Drug injury compensation plans. Am. J. Comp. Law. **1**, 297–323 (1982)
40. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press, Cambridge (2016)
41. Guynn, J.: Google photos labelled black people 'gorillas'. *USA today.* http://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/ (2015). Accessed 15 Mar 2021
42. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020)
43. Hauben, M., Bate, A.: Decision support methods for the detection of adverse events in post-marketing data. Drug Discov. Today **14**(7–8), 343–357 (2009)
44. Herkert, J., Borenstein, J., Miller, K.: The Boeing 737 MAX: lessons for engineering ethics. Sci. Eng. Ethics **26**(6), 2957–2974 (2020)
45. High-Level Expert Group on AI of the EU.: Ethics guidelines for trustworthy AI | Shaping Europe's digital future". https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019). Accessed 15 Mar 2021
46. Hoffmann, A.L.: Terms of inclusion: data, discourse, violence. New Media Soc. **23**(12), 3539–3556 (2020)
47. Hoofnagle, C.J., van der Sloot, B., Borgesius, F.Z.: The European Union general data protection regulation: what it is and what it means. Inf. Commun. Technol. Law **28**(1), 65–98 (2019)
48. Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep learning. Electron. Markets **31**, 685–695 (2021)
49. Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Oxford (2019)
50. Kim, Y.C., Dema, B., Reyes-Sandoval, A.: COVID-19 vaccines: breaking record times to first-in-human trials. NPJ Vacc. **5**(1), 1–3 (2020)
51. Lee, N.T., Resnick, P., Barton, G.: Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms. Brookings Institute, Washington, DC (2019)
52. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)
53. McDuff, D., Cheng, R., Kapoor, A.: Identifying bias in AI using simulation. arXiv:1810.00471 (2018)
54. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
55. Mills, C.W.: The Racial Contract. Cornell University Press, Ithaca (2014)
56. Müller, V.C. (Summer 2021 Edition), Zalta, E.N.: (eds.) Ethics of artificial intelligence and robotics. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/. Accessed 18 Mar 2021
57. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
58. Nabirahni, D.M., Evans, B.R., Persaud, A.: Al-Khwarizmi (algorithm) and the development of algebra. Math. Teach. Res. J. **11**(1–2), 13–17 (2019)
59. Nielsen, M.W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H.J., Joshi, A., Leahey, E., Smith-Doerr, L., Woolley, A.W., Schiebinger, L.: Opinion: gender diversity leads to better science. Proc. Natl. Acad. Sci. **114**(8), 1740–1742 (2017)
60. Noble, S.U.: Algorithms of Oppression. New York University Press, New York (2018)
61. Northpointe Inc.: Measurement & treatment implications of COMPAS core scales. Technical report, Northpointe Inc. https://www.michigan.gov/documents/corrections/Timothy Brenne Ph.D. Meaning and treatment implications of COMPA core scales 297495 7.pdf. Accessed 2 Feb 2020 (2009)
62. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
63. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 13 (2019)
64. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Penguin Books Limited, New York (2016)
65. Onuoha, M.: Notes on Algorithmic Violence. https://github.com/MimiOnuoha/On-Algorithmic-Violence (2018). Accessed 20 Aug 2021
66. Opeyemi, B.: *Deployment of Machine learning Models Demystified (Part 1). Towards Data Science* (2019)
67. Pateman, C.: The Sexual Contract. Wiley, Weinheim (2018)

68. Podesta Report. Exec.: Office of the President, big data: seizing opportunities, preserving values. https://obamawhitehouse.archives.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf (2014). Accessed 15 Aug 2021
69. Reed, C.: How should we regulate artificial intelligence? Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **376**(2128), 20170360 (2018)
70. Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, pp. 1–22. AI Now Institute (2018)
71. Ricardo, B.Y.: Bias on the web. Commun. ACM **61**(6), 54–61 (2018)
72. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, New York (2016)
73. Sandler, R., Basl, J.: Building Data and AI Ethics Committees. North Eastern University Ethics Institute and Accenture. https://cssh.northeastern.edu/informationethics/wp-content/uploads/sites/44/2020/08/811330-AI-Data-Ethics-Committee-Report_V10.0.pdf (2019). Accessed 7 May 2021
74. Santoro, M.A., Gorrie, T.M.: Ethics and the Pharmaceutical Industry. Cambridge University Press, Cambridge (2005)
75. Sax, L.J., Lehman, K.J., Jacobs, J.A., Kanny, M.A., Lim, G., Monje-Paulson, L., Zimmerman, H.B.: Anatomy of an enduring gender gap: the evolution of women's participation in computer science. J. Higher Educ. **88**(2), 258–293 (2017)
76. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T.: Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020)
77. Sedgwick, P.: Phases of clinical trials. BMJ **343**, d6068 (2011)
78. Shapira, R., Zingales, L.: Is Pollution Value-Maximizing? The DuPont case (No. w23866). National Bureau of Economic Research (2017)
79. Shields, M.: Women's participation in Seattle's high-tech economy. https://smartech.gatech.edu/bitstream/handle/1853/53790/madelyn_shields_womens_participation_in_seattles_hightech_economy.pdf (2015). Accessed 15 Aug 2021
80. Spiekermann, S.: Ethical IT innovation: a value-based system design approach. CRC Press, Boca Raton (2015)
81. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. arXiv:1901.10002 (2019)
82. Swift, S.: Gender Disparities in the Tech Industry: The Effects of Gender and Stereotypicability on Perceived Environmental Fit. In: *2015 NCUR* (2015)
83. The National Archives.: Equality Act 2010. [online] https://www.legislation.gov.uk/ukpga/2010/15/contents. Accessed 15 June 2021
84. Thelisson, E., Padh, K., Celis, L.E.: Regulatory mechanisms and algorithms towards trust in AI/ML. In: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), Melbourne, Australia (2017)
85. Tolan, S.: Fair and unbiased algorithmic decision making: current state and future challenges. arXiv:1901.04730 (2019)
86. Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.P., Humbert, M., Juels, A., Lin, H.: FairTest: discovering unwarranted associations in data-driven applications. In: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 401–416. IEEE (2017)
87. US Census Bureau, Bureau of Justice Statistics.: https://data.census.gov/cedsci/table?q=S0201&t=400%20-%20Hispanic%20or%20Latino%20%28of%20any%20race%29%20%28200-299%29%3A451%20-%20White%20alone,%20not%20Hispanic%20or%20Latino%3A453%20-%20Black%20or%20African%20American%20alone,%20not%20Hispanic%20or%20Latino&tid=ACSSPP1Y2019.S0201 (2019). Accessed 22 Apr 2021
88. Van Wel, L., Royakkers, L.: Ethical issues in web data mining. Ethics Inf. Technol. **6**(2), 129–140 (2004)
89. Van Wynsberghe, A., Robbins, S.: Critiquing the reasons for making artificial moral agents. Sci. Eng. Ethics **25**(3), 719–735 (2019)
90. Verdin, J., Funk, C., Senay, G., Choularton, R.: Climate science and famine early warning. Philos. Trans. R. Soc. B Biol. Sci. **360**(1463), 2155–2168 (2005)
91. Vincent, J.: Amazon reportedly scraps internal AI recruiting tool that was biased against women. *The Verge*. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report (2018). Accessed 28 Mar 2021
92. Wajcman, J.: Feminism Confronts Technology. Penn State Press, Pennsylvania (1991)
93. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
94. Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., Jordan, M.I.: Robust optimization for fairness with noisy protected groups. arXiv:2002.09343 (2020)
95. Washington, A.L.: How to argue with an algorithm: lessons from the COMPAS-ProPublica debate. Colo. Tech. LJ **17**, 131 (2018)
96. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S.: Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Nuffield Foundation, London (2019)
97. Whitehouse, G., Diamond, C.: Reproducing gender inequality: segregation and career paths in information technology jobs in Australia. Reworking **1**, 555–564 (2005)
98. Winfield, A.F., Jirotka, M.: The case for an ethical black box. In: Annual Conference Towards Autonomous Robotic Systems, pp. 262–273. Springer, Cham (2017)
99. Woolley, S.C., Howard, P.N. (eds.) Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media. Oxford University Press, Oxford (2018)
100. World Prison Brief.: https://prisonstudies.org/country/united-states-america (2018). Accessed 22 Apr 2021
101. Yasser, Q.R., Al Mamun, A., Ahmed, I.: Corporate social responsibility and gender diversity: insights from Asia Pacific. Corp. Soc. Responsib. Environ. Manag. **24**(3), 210–221 (2017)
102. Zeng, Z.: Jail Inmates in 2018, US Census Bureau, Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/jail-inmates-2018. Accessed 22 June 2021 (2020)
103. Zhou, N., Zhang, Z., Nair, V.N., Singhal, H., Chen, J., Sudjianto, A.: Bias, Fairness, and Accountability with AI and ML Algorithms. arXiv:2105.06558 (2021)
104. Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. United States: PublicAffairs (2019)