AI & SOCIETY (2022) 37:215–230
https://doi.org/10.1007/s00146-021-01154-8

OPEN FORUM

The ethics of algorithms: key problems and solutions

Andreas Tsamados[1] · Nikita Aggarwal[1,3] · Josh Cowls[1,2] · Jessica Morley[1] · Huw Roberts[1] · Mariarosaria Taddeo[1,2] · Luciano Floridi[1,2]

Received: 27 July 2020 / Accepted: 22 January 2021 / Published online: 20 February 2021

Abstract
Research on the ethics of algorithms has grown substantially over the past decade. Alongside the exponential development and application of machine learning algorithms, new ethical problems and solutions relating to their ubiquitous use in society have been proposed. This article builds on a review of the ethics of algorithms published in 2016 (Mittelstadt et al. Big Data Soc 3(2), 2016). The goals are to contribute to the debate on the identification and analysis of the ethical implications of algorithms, to provide an updated analysis of epistemic and normative concerns, and to offer actionable guidance for the governance of the design, development and deployment of algorithms.

Keywords Algorithm · Artificial intelligence · Autonomy · Digital ethics · Explainability · Fairness · Machine learning · Privacy · Responsibility · Transparency · Trust

1 Introduction

Algorithms have become a key element underpinning crucial services and infrastructures of information societies. Individuals interact with recommender systems—algorithmic systems that make suggestions about what a user may like—on a daily basis, be it to choose a song, a movie, a product or even a friend (Paraschakis 2017; Perra and Rocha 2019; Milano et al. 2020). At the same time, schools and hospitals (Obermeyer et al. 2019; Zhou et al. 2019; Morley et al. 2019a, b), financial institutions (Lee and Floridi 2020; Aggarwal 2020) courts (Green and Chen 2019; Yu and Du 2019), local governmental bodies (Eubanks 2017; Lewis 2019), and national governments (Labati et al. 2016; Hauer 2019; Taddeo and Floridi 2018a; Taddeo et al. 2019; Roberts et al. 2019), all increasingly rely on algorithms to make significant decisions.

The potential for algorithms to improve individual and social welfare comes with significant ethical risks (Floridi and Taddeo 2016). Algorithms are not ethically neutral. Consider, for example, how the outputs of translation and search engine algorithms are largely perceived as objective, yet frequently encode language in gendered ways (Larson 2017; Prates et al. 2019). Bias has also been reported in algorithmic advertisement, with opportunities for higher-paying jobs and jobs within the field of science and technology advertised to men more often than to women (Datta et al. 2015; Lambrecht and Tucker 2019). Likewise, prediction algorithms used to manage the health data of millions of patients in the United States exacerbate existing problems, with white patients given measurably better care than comparably similar, black patients (Obermeyer et al. 2019). While solutions to these issues are being discussed and designed, the number of algorithmic systems exhibiting ethical problems continues to grow.

Since 2012, artificial intelligence (AI) has been experiencing a new 'summer', both in terms of the technical advances being made and the attention that the field has received from academics, policy makers, technologists, and investors

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[2] Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

[3] Faculty of Law, University of Oxford, St. Cross Building, St. Cross Road, Oxford OX1 3UL, UK

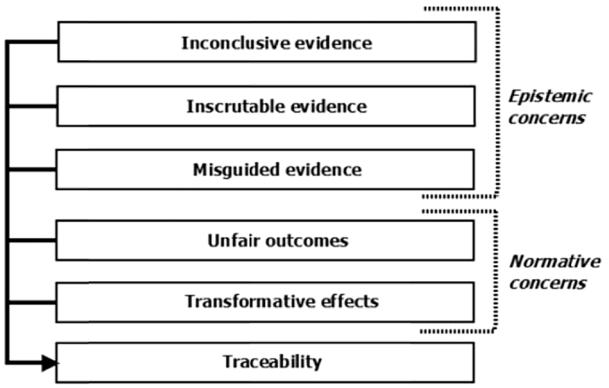

Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

(Perrault et al. 2019). Within this, there has been a growing body of research on the ethical implications of algorithms, particularly in relation to fairness, accountability, and transparency (Lee 2018; Hoffmann et al. 2018; Shin and Park 2019). In 2016, our research group at the Digital Ethics Lab published a comprehensive study that sought to map these ethical concerns (Mittelstadt et al. 2016). However, this is a fast-changing field and both novel ethical problems and ways to address them have emerged, making it necessary to improve and update that study.

In particular, work on the ethics of algorithms has increased significantly since 2016, when national governments, non-governmental organisations, and private companies started to take a prominent role in the conversation on "fair" and "ethical" AI and algorithms (Sandvig et al. 2016; Binns 2018; Selbst et al. 2019; Wong 2019; Ochigame 2019). Both the quantity and the quality of the research available on the topic have expanded enormously. Given these changes, this article updates our previous work in light of new insights into the ethics of algorithms, updates the initial analysis, includes references to the literature that were missed by the original review, and extends the analysed topics, including for example work on AI for social good (see the Sect. 9). At the same time, the conceptual map proposed in 2016 (see Fig. 1) remains a fruitful framework for reviewing the current debate on the ethics of algorithms, identifying the ethical problems that algorithms give rise to, and the solutions that have been proposed in recent relevant literature. Specifically, in Sect. 2, we summarise the conceptual map and detail our methodology for the literature review. In Sects. 3–8, we offer systematic search and review (in the methodological sense specified by Grant and Booth 2009) on the ethics of algorithms and draw links with the types of ethical concerns previously identified. Section 9 concludes the article with an overview.

2 Map of the ethics of algorithms

There is little agreement in the relevant literature on the definition of an algorithm. The term is often used to indicate both the formal definition of an algorithm as a mathematical construct, with ''a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions" (Hill 2016, 47), as well as domain-specific understandings which focus on the implementation of these mathematical constructs into a technology configured for a specific task. In this article, we decided to maintain the same approach adopted in the 2016 article and to focus on the ethical issues posed by algorithms as mathematical constructs, their implementations as programs and configurations (applications), and the ways in which these can be addressed. We consider algorithms that are used to (1) turn data into evidence for a given outcome, which is used to, (2) trigger and motivate an action that may have ethical consequences. Actions (1) and (2) may be performed by (semi-)autonomous algorithms—such as machine learning (ML) algorithms—and this complicates, (3) the attribution of responsibility for the effects of actions that an algorithm may trigger. Here, ML is of particular interest, as a field which includes deep learning architectures. Computer systems deploying ML algorithms may be described as "autonomous" or "semi-autonomous", to the extent that their outputs are induced from data and thus, non-deterministic.

Based on this approach, we used the conceptual map shown in Fig. 1 to identify the ethical issues that algorithms pose. The map identifies six ethical concerns, which define the conceptual space of the ethics of algorithms as a field of research. Three of the ethical concerns refer to epistemic factors, specifically: inconclusive, inscrutable, and misguided evidence. Two are explicitly normative: unfair outcomes and transformative effects; while one—traceability—is relevant both for epistemic and normative purposes.

The epistemic factors in the map highlight the relevance of the quality and accuracy of the data for the justifiability of the conclusions that algorithms reach and which, in turn, may shape morally-loaded decisions affecting individuals, societies, and the environment. The normative concerns identified in the map refer explicitly to the ethical impact of algorithmically-driven actions and decisions, including lack of transparency (opacity) of algorithmic processes, unfair outcomes, and unintended consequences. Epistemic and normative concerns, together with the distribution of the design, development, and deployment of algorithms make it hard to trace the chain of events and factors leading to a given outcome, thus, hindering the possibility of identifying its cause, and of attributing moral responsibility for it. This is what the sixth ethical concern, traceability, refers to.

Table 1 Systematic literature search results

Database	Keywords	Returned
Scopus	Algorithm* AND ethics	1682
	AND traceability	552
	AND fairness	2890
	AND autonomy	1190
	AND (accountability OR responsibility)	1205
	AND (transparency OR scrutability OR opacity)	1567
	AND discrimination	3369
Web of Science	Algorithm* AND ethics	1310
	AND traceability	700
	AND fairness	2604
	AND autonomy	1205
	AND (accountability OR responsibility)	1503
	AND (transparency OR scrutability OR opacity)	2788
	AND discrimination	4730
Philpapers	Algorithm* AND ethics	370
Google Scholar	Algorithm* AND ethics	100[a]

[a]About 93 000 returned, first 100 reviewed

It is important to stress that this conceptual map can be interpreted at both a micro- and macro-ethical level. At the micro-ethical level, it sheds light on the ethical problems that particular algorithms may pose. By highlighting how these issues are inseparable from those related to data and responsibilities, it shows the need to take a macro-ethical approach to addressing the ethics of algorithms as part of a wider conceptual space, namely, digital ethics (Floridi and Taddeo 2016). As Floridi and Taddeo argue:

> "While they are distinct lines of research, the ethics of data, algorithms and practices are obviously intertwined … [Digital] ethics must address the whole conceptual space and hence all three axes of research together, even if with different priorities and focus" (Floridi and Taddeo 2016, 4).

In the remainder of this article, we address each of these six ethical concerns in turn, offering an updated analysis of the ethics of algorithms literature (at a micro level), with the goal of contributing to the debate on digital ethics (at a macro level).

A systematic literature search was performed via keyword queries on four widely used reference repositories to identify and analyse the literature on the ethics of algorithms (seeTable 1). Four keywords were used to describe an algorithm: 'algorithm', 'machine learning', 'software' and 'computer program'.[1] The search was limited to publications made available between November 2016 and March 2020.

The search identified 4891 unique papers for review.[2] After initial review of title/abstract, 180 papers were selected for a full review. Of these, 62 were rejected as off-topic, leaving 118 articles for a full review. There are all listed in the reference list of the paper. Another 37 articles and books were reviewed and referenced in this paper to provide additional information regarding specific ethical issues and solutions (eg. technical details, examples and tools). These were sourced from the bibliographies of the 118 articles we reviewed as well as provided on an ad-hoc basis when agreed upon by the authors as being helpful for clarification.

3 Inconclusive evidence leading to unjustified actions

Research focusing on inconclusive evidence refers to the way in which non-deterministic, ML algorithms produce outputs that are expressed in probabilistic terms (James et al. 2013; Valiant 1984). These types of algorithms generally identify association and correlation between variables in the underlying data, but not causal connections. As such, they encourage the practice of *apophenia*: "seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions" (boyd and Crawford 2012, 668). This is highly problematic, as patterns identified by algorithms may be the result of inherent properties of the system modelled by the data,

[1] The literature search was limited to English language articles in peer-reviewed journals and conference proceedings.

[2] Many of which were purely technical in nature, especially for "discrimination" and "(transparency OR scrutability OR opacity)".

of the datasets (that is, of the model itself, rather than the underlying system), or of skillful manipulation of datasets (properties neither of the model nor of the system). This is the case, for example, of Simpson's paradox, when trends that are observed in different groups of data reverse when the data is aggregated (Blyth 1972). In the last two cases, poor quality of the data leads to inconclusive evidence to support human decisions.

Recent research has underlined the concern that inconclusive evidence can give rise to serious ethical risks. For example, focusing on non-causal indicators may distract attention from the underlying causes of a given problem (Floridi et al. 2020). Even with the use of causal methods, the available data may not always contain enough information to justify an action or make a decision fair (Olhede and Wolfe 2018, 7). Data quality—the timeliness, completeness and correctness of a dataset—constrains the questions that can be answered using a given dataset (Olteanu et al. 2016). Additionally, the insights that can be extracted from datasets are fundamentally dependent on the assumptions that guided the data collection process itself (Diakopoulos and Koliska 2017). For example, algorithms designed to predict patient outcomes in clinical settings rely entirely on data inputs that can be quantified (e.g. vital signs and previous success rates of comparative treatments), whilst ignoring other emotional facts (e.g. the willingness to live) which can have a significant impact on patient outcomes, and thus, undermine the accuracy of the algorithmic prediction (Buhmann, Paßmann, and Fieseler 2019). This example highlights how insights stemming from algorithmic data processing can be uncertain, incomplete, and time-sensitive (Diakopoulos and Koliska 2017).

One may embrace a naïve, inductivist approach and assume that inconclusive evidence can be avoided if algorithms are fed enough data, even if a causal explanation for these results cannot be established. Yet, recent research rejects this view. In particular, literature focusing on the ethical risks of racial profiling using algorithmic systems has demonstrated the limits of this approach highlighting, among other things, that long-standing structural inequalities are often deeply embedded in the algorithms' datasets and are rarely, if ever, corrected for (Hu 2017; Turner Lee 2018; Noble 2018; Benjamin 2019; Richardson et al. 2019; Abebe et al. 2020). More data by themselves do not lead to greater accuracy or greater representation. On the contrary, they may exacerbate issues of inconclusive data by enabling correlations to be found where there really are none. As Ruha Benjamin (2020) put it "computational depth without historical or sociological depth is just superficial learning [not deep learning]". These limitations pose serious constraints on the justifiability of algorithmic outputs, which could have a negative impact on individuals or an entire population due to suboptimal inferences or, in the case of the physical sciences, even tip the evidence for or against "a specific scientific theory" (Ras et al. 2018, 10). This is why it is crucial to ensure that data fed to algorithms are validated independently, and data retention and reproducibility measures are in place to mitigate inconclusive evidence leading to unjustified actions, along with auditing processes to identify unfair outcomes and unintended consequences (Henderson et al. 2018; Rahwan 2018; Davis and Marcus 2019; Brundage et al. 2020).

The danger arising from inconclusive evidence and erroneous actionable insights also stems from the perceived mechanistic objectivity associated with computer-generated analytics (Karppi 2018; Lee 2018; Buhmann et al. 2019). This can lead to human decision-makers ignoring their own experienced assessments—so-called 'automation bias' (Cummings 2012)—or even shirking part of their responsibility for decisions (see Traceability below) (Grote and Berens 2020). As we shall see in Sects. 4 and 8, a lack of understanding of how algorithms generate outputs exacerbates this problem.

4 Inscrutable evidence leading to opacity

Inscrutable evidence focuses on problems related to the lack of transparency that often characterise algorithms (particularly ML algorithms and models); the socio-technical infrastructure in which they exist; and the decisions they support. Lack of transparency—whether inherent due to the limits of technology or acquired by design decisions and obfuscation of the underlying data (Lepri et al. 2018; Dahl 2018; Ananny and Crawford 2018; Weller 2019)—often translates into a lack of scrutiny and/or accountability (Oswald 2018; Fink 2018; Webb et al. 2019) and leads to a lack of "trustworthiness" (see Al-Hleg 2019).

According to the recent literature, factors contributing to the overall lack of algorithmic transparency include the cognitive impossibility for humans to interpret massive algorithmic models and datasets; a lack of appropriate tools to visualise and track large volumes of code and data; code and data that are so poorly structured that they are impossible to read; and ongoing updates and human influence over a model (Diakopoulos and Koliska 2017; Stilgoe 2018; Zerilli et al. 2019; Buhmann et al. 2019). Lack of transparency is also an inherent characteristic of self-learning algorithms, which alter their decision logic (produce new sets of rules) during the learning process, making it difficult for developers to maintain a detailed understanding of why certain changes were made (Burrell 2016; Buhmann et al. 2019). However, this does not necessarily translate into opaque outcomes, as even without understanding each logical step, developers can adjust hyperparameters, the parameters that govern the training process, to test for various outputs. In this respect, Martin (2019) stresses that, while the difficulty of explaining ML algorithms' outputs is certainly real, it is important not to let this difficulty incentivise organisations to develop complex systems to shirk responsibility.

Lack of transparency can also result from the malleability of algorithms, whereby algorithms can be reprogrammed in a continuous, distributed, and dynamic way (Sandvig et al. 2016). Algorithmic malleability allows developers to monitor and improve an already-deployed algorithm, but it may also be abused to blur the history of its evolution and leave end-users in a state of confusion about the affordances of a given algorithm (Ananny and Crawford 2018). Consider for example Google's main search algorithm. Its malleability enables the company to make continuous revisions, suggesting a permanent state of destabilisation (Sandvig et al. 2016). This requires those affected by the algorithm to monitor it constantly and update their understanding accordingly –an impossible task for most (Ananny and Crawford 2018).

As Floridi and Turilli (2009, 105) note, transparency is not an "ethical principle in itself but a pro-ethical condition for enabling or impairing other ethical practices or principles". And indeed, complete transparency can itself cause distinct ethical problems (Ananny and Crawford 2018): transparency can provide users with some critical information about the features and limitations of an algorithm, but it can also overwhelm users with information and thus, render the algorithm more opaque (Kizilcec 2016; Ananny and Crawford 2018). Other research stress that excessive focus on transparency can be detrimental to innovation and unnecessarily divert resources that could instead be used to improving safety, performance and accuracy (Danks and London 2017; Oswald 2018; Ananny and Crawford 2018; Weller 2019). For example, the debate over prioritising transparency (and explainability) is especially contentious in the context of medical algorithms (Robbins 2019).

Transparency can enable individuals to game the system (Martin 2019; Magalhães 2018; Floridi et al. 2020). Knowledge about the source of a dataset, the assumptions under which sampling was done, or the metrics that an algorithm uses to sort new inputs, may be used to figure out ways to take advantage of an algorithm (Szegedy et al. 2014; Yampolskiy 2018). Yet, the ability to game algorithms is only within reach for some groups of the population—those with higher digital literacy for example—thus, creating another form of social inequality (Martin 2019; Bambauer and Zarsky 2018). Therefore, confusing transparency for an end in itself, instead of a pro-ethical factor (Floridi 2017) enabling crucial ethical practices, may not solve existing ethical problems related to the use of algorithms and, indeed, pose new ones. This is why it is important to distinguish between the different factors that may hinder transparency of algorithms, identify their cause, and nuance the call for transparency by specifying which factors are required and at which layers of algorithmic systems they should be addressed (Diakopoulos and Koliska 2017).

There are different ways of addressing the problems related to lack of transparency. For example, Gebru et al. propose that the constraints on transparency posed by the malleability of algorithms can be addressed, in part, by using standard documentary procedures similar to those deployed in the electronics industry, where.

> "every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information" (Gebru et al. 2020, 2).

Unfortunately, publicly available documentation is currently uncommon in the development of algorithmic systems and there is no agreed-upon format for what should be included when documenting the origin of a dataset (Arnold et al. 2019; Gebru et al. 2020).

Although relatively nascent, another potentially promising approach to enforcing algorithmic transparency is the use of technical tools to test and audit algorithmic systems and decision-making. Testing whether algorithms exhibit negative tendencies, like unfair discrimination, and auditing a prediction or decision trail in detail, can help maintain a high level of transparency (Weller 2019; Malhotra et al. 2018; Brundage et al. 2020). To this end, discursive frameworks have been developed to help businesses and public sector organisations understand the potential impacts of opaque algorithms, thus encouraging good practices (ICO 2020). For instance, the AI Now Institute at New York University has produced algorithmic impact assessment guidance, which seeks to raise awareness and improve dialogue over potential harms of ML algorithms (Reisman et al. 2018). This includes the two aims of enabling developers to design more transparent, and therefore more trustworthy ML algorithms, and of improving the public understanding and control of algorithms. In the same vein, Diakopoulos and Koliska have provided a comprehensive list of "transparency factors" across four layers of algorithmic systems: data, model, inference, and interface. Factors include, inter alia.

> "uncertainty (e.g. error margins), timeliness (e.g. when was the data collected), completeness or missing elements, sampling method, provenance (e.g. sources), and volume (e.g. of training data used in machine learning)" (Diakopoulos and Koliska 2017, 818).

Effective transparency procedures are likely, and indeed ought to, involve an interpretable explanation of the internal processes of these systems. Buhmann et al. (2019) argue that while a lack of transparency is an inherent feature of many ML algorithms, this does not mean that improvements cannot be made (Watson et al. 2019). For example, companies like Google and IBM have increased their efforts to make ML algorithms more interpretable and inclusive by making tools such as Explainable AI, AI Explainability 360, and the What-If Tool publicly available. These tools provide developers and also the general public with interactive visual interfaces that

improve human readability, explore various model results, provide case-based reasoning, directly interpretable rules, and even identify and mitigate unwanted biases in datasets and algorithmic models (Mojsilovic 2018; Wexler 2018).

However, explanations for ML algorithms are constrained by the type of explanation sought, the fact that decisions are often multi-dimensional in their nature, and that different users may require different explanations (Edwards and Veale 2017). Identifying appropriate methods for providing explanations has been a problem since the late 1990s (Tickle et al. 1998), but contemporary efforts can be categorised into two main approaches: subject-centric explanations and model-centric explanations (Doshi-Velez and Kim 2017; Lee et al. 2017; Baumer 2017; Buhmann et al. 2019). In the former, the accuracy and length of the explanation is tailored to users and their specific interactions with a given algorithm (see for example [Green and Viljoen 2020] and the game-like model proposed by [Watson and Floridi 2020]); in the latter, explanations concern the model as a whole and do not depend on their audience.

Explainability is particularly important when considering the rapidly growing number of open source and easy-to-use models and datasets. Increasingly, non-experts are experimenting with state-of-the-art algorithmic models widely available via online libraries or platforms, like GitHub, without always fully grasping their limits and properties (Hutson 2019). This has prompted scholars to suggest that, to tackle the issue of technical complexity, it is necessary to invest more heavily in public education to enhance computational and data literacy (Lepri et al. 2018). Doing so would seem to be an appropriate long-term solution to the multi-layered issues introduced by ubiquitous algorithms, and open-source software is often cited as critical to the solution (Lepri et al. 2018).

5 Misguided evidence leading to unwanted bias

Developers are predominantly focused on ensuring that their algorithms perform the tasks for which they were designed. Thus, the type of thinking that guides developers is essential to understanding the emergence of bias in algorithms and algorithmic decision-making. Some scholars refer to the dominant thinking in the field of algorithm development as being defined by "algorithmic formalism"—an adherence to prescribed rules and form (Green and Viljoen 2020, 21). While this approach is useful for abstracting and defining analytical processes, it tends to ignore the social complexity of the real world (Katell et al. 2020). Indeed, this approach leads to algorithmic interventions that strive to be 'neutral' but in doing so, it risks entrenching existing social conditions (Green and Viljoen 2020, 20), while creating the illusion of precision (Karppi 2018; Selbst et al. 2019). For these reasons, the use of algorithms in some settings is questioned altogether (Selbst et al. 2019; Mayson 2019; Katell et al. 2020; Abebe et al. 2020). For example, a growing number of scholars criticise the use of algorithm-based risk assessment tools in court settings (Berk et al. 2018; Abebe et al. 2020).

Some scholars affirm the limits of abstractions with regard to unwanted bias in algorithms and argue for the need to develop a sociotechnical frame to address and improve the fairness of algorithms (Edwards and Veale 2017; Selbst et al. 2019; Wong 2019; Katell et al. 2020; Abebe et al. 2020). In this respect, Selbst et al. (2019, 60–63) point to five abstraction "traps", or failures to account for the social context in which algorithms operate, which persist in algorithmic design due to the absence of a sociotechnical frame, namely:

1. A failure to model the entire system over which a social criterion, such as fairness, will be enforced;
2. A failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context;
3. A failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms;
4. A failure to understand how the insertion of technology into an existing social system changes the behaviours and embedded values of the pre-existing system; and
5. A failure to recognize the possibility that the best solution to a problem may not involve technology.

The term 'bias' often comes with a negative connotation, but it is used here to denote a "deviation from a standard" (Danks and London 2017, 4692), which can occur at any stage of the design, development, and deployment process. The data used to train an algorithm is one of the main sources from which bias emerges (Shah 2018), through preferentially sampled data or from data reflecting existing societal bias (Diakopoulos and Koliska 2017; Danks and London 2017; Binns 2018; Malhotra et al. 2018). For example, morally problematic structural inequalities that disadvantage certain ethnicities may not be apparent in data and thus not corrected for (Nobles 2018; Benjamin 2019). Additionally, data used to train algorithms are seldom obtained "according to any specific experimental design" (Olhede and Wolfe 2018, 3) and are used even though they may be inaccurate, skewed, or systemically biased, offering a poor representation of a population under study (Richardson et al. 2019).

One possible approach to mitigating this problem is to exclude intentionally some specific data variables from informing algorithmic decision-making. Indeed, the

processing of statistically relevant sensitive or "protected variables"—such as gender or race—is typically limited or prohibited under anti-discrimination and data protection law, to limit the risks of unfair discrimination. Unfortunately, even if protections for specific classes can be encoded in an algorithm, there could always be biases that were not considered *ex ante*, as in the case, for example, of language models reproducing heavily male-focused texts (Fuster et al. 2017; Doshi-Velez and Kim 2017). Even while bias may be anticipated and protected variables excluded from the data, unanticipated proxies for these variables could still be used to reconstruct biases, leading to "bias by proxy" that is difficult to detect and avoid (Fuster et al. 2017; Gillis and Spiess 2019).

At the same time, there may be good reasons to rely on statistically biased estimators in algorithmic processing, as they can be used to mitigate training data bias. In this way, one type of problematic algorithmic bias is counterbalanced by another type of algorithmic bias or by introducing compensatory bias when interpreting algorithmic outputs (Danks and London 2017). Simpler approaches to mitigating bias in data involve piloting algorithms in different contexts and with various datasets (Shah 2018). Having a model, its datasets, and metadata (on provenance) published to enable external scrutiny can also help correct unseen or unwanted bias (Shah 2018). It is also worth noting that so-called 'synthetic data', or algorithmically generated data, produced via reinforcement learning or generative adversarial networks (GANs) offer an opportunity to address certain issues of data bias (Floridi 2019a; Xu et al. 2018). Fair data generation with GANs may help diversify datasets used in computer vision algorithms (Xu et al. 2018). For example, StyleGAN2 (Karras et al. 2019) is able to produce high-quality images of non-existing human faces and has proven to be especially useful in creating diverse datasets of human faces, something that many algorithmic systems for facial recognition currently lack (Obermeyer et al. 2019; Kortylewski et al. 2019; Harwell 2020).

Unwanted bias also occurs due to improper deployment of an algorithm. Consider transfer context bias: the problematic bias that emerges when a functioning algorithm is used in a new environment. For example, if a research hospital's healthcare algorithm is used in a rural clinic and assumes that the same level of resources are available to the rural clinic as the research hospital, the healthcare resource allocation decisions generated by the algorithm will be inaccurate and flawed (Danks and London 2017).

In the same vein, Grgić-Hlača et al. (2018) warn of vicious cycles when algorithms make misguided chain assessments. For example, in the context of the COMPAS risk-assessment algorithm, one of the assessment criteria for predicting recidivism is the criminal history of a defendant's friends. It follows that having friends with a criminal history would create a vicious cycle in which a defendant with convicted friends will be deemed more likely to offend, and therefore sentenced to prison, hence increasing the number of people with criminal records in a given group on the basis of mere correlation (Grgić-Hlača et al. 2018; Richardson et al. 2019).

High-profile examples of algorithmic bias in recent years—not least investigative reporting around the COMPAS system (Angwin et al. 2016)—have led to a growing focus on issues of algorithmic fairness. The definition and operationalisation of algorithmic fairness have become "urgent tasks in academia and industry" (Shin and Park 2019), as the significant uptick in the number of papers, workshops and conferences dedicated to 'fairness, accountability and transparency' (FAT) highlights (Hoffmann et al. 2018; Ekstrand and Levy 2018; Shin and Park 2019). We analyse key topics and contributions in this area in the next section.

6 Unfair outcomes leading to discrimination

There is widespread agreement on the need for algorithmic fairness, particularly to mitigate the risks of direct and indirect discrimination (under US law, 'disparate treatment' and 'disparate impact', respectively) due to algorithmic decisions (Barocas and Selbst 2016; Grgić-Hlača et al. 2018; Green and Chen 2019). Yet there remains a lack of agreement among researchers on the definition, measurements and standards of algorithmic fairness (Gajane and Pechenizkiy 2018; Saxena et al. 2019; Lee 2018; Milano et al. 2020). Wong (2019) identifies up to 21 definitions of fairness across the literature and such definitions are often mutually inconsistent (Doshi-Velez and Kim 2017).

There are many nuances in the definition, measurement, and application of different standards of algorithmic fairness. For instance, algorithmic fairness can be defined both in relation to groups as well as individuals (Doshi-Velez and Kim 2017). Four main definitions of algorithmic fairness have gained prominence in the recent literature (see for example [Kleinberg et al. 2016; Corbett-Davies and Goel 2018]):

1. Anti-classification, which refers to protected categories, such as race and gender, and their proxies not being explicitly used in decision making;
2. Classification parity, which regards a model as being fair if common measures of predictive performance, including false positive and negative rates, are equal across protected groups;
3. Calibration, which considers fairness as a measure of how well-calibrated an algorithm is between protected groups;

4. Statistical parity, which defines fairness as an equal average probability estimate over all members of protected groups.

However, each of these commonly used definitions of fairness has drawbacks and are generally mutually incompatible (Kleinberg et al. 2016). Taking anti-classification as an example, protected characteristics, such as race, gender and religion, cannot simply be removed from training data to prevent discrimination, as noted above (Gillis and Spiess 2019). Structural inequalities mean that formally non-discriminatory data points such as postcodes can act as proxies for, and be used, either intentionally or unintentionally, to infer protected characteristics, like race (Edwards and Veale 2017).

There are important cases where it is appropriate to consider protected characteristics to make equitable decisions. For example, lower female reoffending rates mean that excluding gender as an input in recidivism algorithms would leave women with disproportionately high-risk ratings (Corbett-Davies and Goel 2018). Because of this, Binns (2018) stresses the importance of considering the historical and sociological context that cannot be captured in the data presented to algorithms but that can inform contextually appropriate approaches to fairness in algorithms. It is also critical to note that algorithmic models can often produce unexpected outcomes, contrary to human intuitions and perturb their understanding. For example, as Grgić-Hlača et al. (2018) highlight, using features that people believe to be fair can in some cases increase the racism exhibited by algorithms and decrease accuracy.

Regarding methods for improving algorithmic fairness, Veale and Binns (2017) and Katell et al. (2020) offer two approaches. The first envisages a third-party intervention, whereby an entity external to the provider of algorithms would hold data on sensitive or protected characteristics and attempt to identify and reduce discrimination caused by the data and models. The second approach proposes a collaborative knowledge-based method which would focus on community-driven data resources containing practical experiences of ML and modelling (Veale and Binns 2017; Katell et al. 2020). The two approaches are not mutually exclusive, they may bring different benefits depending on contexts of application, and their combination may also be beneficial.

Given the significant impact that algorithmic decisions have on people's lives and the importance of context for choosing appropriate measures of fairness, it is surprising that there has been little effort to capture public views on algorithmic fairness (Lee et al. 2017; Saxena et al. 2019; Binns 2018). Examining public perceptions of different definitions of algorithmic fairness, Saxena et al. (2019, 3) note that in the context of loan decisions people exhibit a preference for a "calibrated fairness definition", or merit-based selection, as compared to "treating similar people similarly" and argue in favour of the principle of affirmative action. In a similar study, Lee (2018) offers evidence suggesting that, when considering tasks that require uniquely human skills, people consider algorithmic decisions to be less fair and algorithms to be less trustworthy.

Reporting on empirical work conducted on algorithmic interpretability and transparency, Webb et al. (2019) reveal that moral references, particularly on fairness, are consistent across participants discussing their preferences on algorithms. The study notes that people tend to go beyond personal preferences to focus instead on "right and wrong behaviour", as a way to indicate the need to understand the context of deployment of the algorithm and the difficulty of understanding the algorithm and its consequences (Webb et al. 2019). In the context of recommender systems, Burke (2017) proposes a multi-stakeholder and multi-sided approach to defining fairness, moving beyond user-centric definitions to include the interests of other system stakeholders.

It has become clear that understanding the public view on algorithmic fairness would help technologists in developing algorithms with fairness principles that align with the sentiments of the general public on prevailing notions of fairness (Saxena et al. 2019, 1). Grounding the design decisions of the providers of an algorithm "with reasons that are acceptable by the most adversely affected" as well as being "open to adjustments in light of new reasons" (Wong 2019, 15) is crucial to improving the social impact of algorithms. It is important to appreciate, however, that measures of fairness are often completely inadequate when they seek to validate models that are deployed on groups of people that are already disadvantaged in society because of their origin, income level, or sexual orientation. We simply cannot "optimise around" (Benjamin 2019) existing economic, social, and political power dynamics (Winner 1980; Benjamin 2019).

7 Transformative effects leading to challenges for autonomy and informational privacy

The collective impact of algorithms has spurred discussions on the autonomy afforded to end users. (Ananny and Crawford 2018; Beer 2017; Taddeo and Floridi 2018b; Möller et al. 2018; Malhotra et al. 2018; Shin and Park 2019; Hauer 2019; Bauer and Dubljević 2020). Algorithm-based services are increasingly featured "within an ecosystem of complex, socio-technical issues" (Shin and Park 2019), which can hinder the autonomy of users. Limits to users' autonomy stem from three sources:

1. pervasive distribution and proactivity of (learning) algorithms to inform users' choice (Yang et al. 2018; Taddeo and Floridi 2018b);
2. users' limited understanding of algorithms;
3. lack of second-order power (or appeals) over algorithmic outcomes (Rubel et al. 2019).

In considering the ethical challenges of AI, Yang et al. (2018, 11) focus on the impact of autonomous, self-learning algorithms on human self-determination and stress that "AI's predictive power and relentless nudging, even if unintentional, should foster and not undermine human dignity and self-determination".

The risks that algorithmic systems may hinder human autonomy by shaping users' choices has been widely reported in the literature and has taken centre stage in most of the high-level ethical principles for AI, including, *inter alia*, those of the European Commission's European Group on Ethics in Science and Technologies, and the UK's House of Lords Artificial Intelligence Committee (Floridi and Cowls 2019). In their analysis of these high-level principles, Floridi and Cowls (2019) note that it does not suffice that algorithms promote people's autonomy: rather, the autonomy of algorithms should be constrained and reversible. Looking beyond the West, the Beijing AI Principles—developed by a consortium of China's leading companies and universities for guiding AI research and development—also emphasise that human autonomy should be respected (Roberts et al. 2020).

Human autonomy can also be limited by the inability of an individual to understand some information or make the appropriate decisions. As Shin and Park suggest, algorithms "do not have the affordance that would allow users to understand them or how best to utilize them to achieve their goals" (Shin and Park 2019, 279). As such, a key issue identified in debates over users' autonomy is the difficulty of striking an appropriate balance between people's own decision-making and that which they delegate to algorithms (Floridi et al. 2018). This is further complicated by a lack of transparency over the decision-making process by which particular decisions are delegated to algorithms. Ananny and Crawford (2018) note that often this process does not account for all stakeholders and is not void of structural inequalities.

As a method of Responsible Research and Innovation (RRI), 'participatory design' is often mentioned for its focus on the design of algorithms to promote the values of end users and protect their autonomy (Whitman et al. 2018; Katell et al. 2020). Participatory design aims at "bringing participants' tacit knowledge and embodied experience into the design process" (Whitman et al. 2018, 2). For example, Rahwan's 'Society-in-the-Loop' (2018) conceptual framework seeks to enable different stakeholders in society to design algorithmic systems before deployment and to amend and reverse the decisions of algorithmic systems that already underlie social activities. This framework aims to maintain a well-functioning "algorithmic social contract", defined as "a pact between various human stakeholders, mediated by machines" (Rahwan 2018, 1). It accomplishes this by identifying and negotiating the values of different stakeholders affected by algorithmic systems as the basis for monitoring adherence to the social contract.

Informational privacy is intimately linked with user autonomy (Cohen 2000; Rössler 2015). Informational privacy guarantees peoples' freedom to think, communicate, and form relationships, among other essential human activities (Rachels 1975; Allen 2011). However, people's increasing interaction with algorithmic systems has effectively reduced their ability to control who has access to information that concerns them and what is being done with it. The vast amounts of sensitive data required in algorithmic profiling and predictions, central to recommender systems, pose multiple issues regarding individuals' informational privacy.

Algorithmic profiling takes place over an indefinite period of time, in which individuals are categorised according to a system's internal logic, and their profiles are updated as new information is obtained about them. This information is typically obtained directly, from when a person interacts with a given system, or indirectly, inferred from algorithmically assembled groups of individuals (Paraschakis 2018). Indeed, algorithmic profiling will also rely on information gathered about other individuals and groups of people that have been categorised in a similar manner to a targeted person. This includes information ranging from characteristics like geographical location and age to information on specific behaviour and preferences, including what type of content a person is likely to seek the most on a given platform (Chakraborty et al. 2019). While this poses a problem of *inconclusive evidence*, it also indicates that if group privacy (Taylor et al. 2017) is not ensured, it may be impossible for individuals to ever remove themselves from the process of algorithmic profiling and predictions (Milano et al. 2020). In other words, individuals' informational privacy cannot be secured without securing group privacy.

Users may not always be aware of, or may not have the ability to gain awareness about, the type of information that is being held about them and what that information is used for. Considering that recommender systems contribute to the dynamic construction of individuals' identities by intervening in their choices, a lack of control over one's information translates in a loss of autonomy.

Giving individuals the ability to contribute to the design of a recommender system can help create more accurate profiles that account for attributes and social categories that would have otherwise not been included in the labelling used by the system to categorise users (Milano et al. 2020).

While the desirability of improving algorithmic profiling will vary with the context, improving the algorithmic design by including feedback from the various stakeholders of the algorithm falls in line with the aforementioned scholarship on RRI and improves users' ability for self-determination (Whitman et al. 2018).

Knowledge about who owns one's data and what is done with them can also help inform trade-offs between informational privacy and information-processing benefits (Sloan and Warner 2018, 21). For example, in medical contexts, individuals are more likely to be willing to share information that can help inform their, or others' diagnostics, less so in the context of job recruitment. Information coordination norms, as Sloan and Warner (2018) argue, can serve to ensure that these trade-offs adapt correctly to different contexts and do not place an excessive amount of responsibility and effort on single individuals. For example, personal information ought to flow differently in the context of law enforcement procedures as compared to a job recruitment process. The European Union's General Data Protection Regulation has played an important role in instituting the basis of such norms (Sloan and Warner 2018).

Finally, a growing scholarship on differential privacy is providing new privacy protection methods for organisations looking to protect their users' privacy while also keeping good model quality, as well as manageable software costs and complexity, striking a balance between utility and privacy (Abadi et al. 2016; Wang et al. 2017; Xian et al. 2017). Technical advancements of this kind, which allow organisations to share publicly a dataset while keeping information about individuals secret (preventing re-identification), and can ensure provable privacy protection on sensitive data, such as genomic data (Wang et al. 2017). Indeed, differential privacy was recently used by Social Science One and Facebook to release safely one of the largest datasets (38 million URLs shared publicly on Facebook) for academic research on the societal impacts of social media (King and Persily 2020).

8 Traceability leading to moral responsibility

The technical limitations of various ML algorithms, such as lack of transparency and lack of explainability, undermine their scrutability and highlight the need for novel approaches to tracing moral responsibility and accountability for the actions performed by ML algorithms. Regarding moral responsibility, Reddy et al. (2019) note a common blurring between technical limitations of algorithms and the broader legal, ethical, and institutional boundaries in which they operate. Even for non-learning algorithms, traditional, linear conceptions of responsibility prove to offer limited guidance in contemporary sociotechnical contexts. Wider sociotechnical structures make it difficult to trace back responsibility for actions performed by distributed, hybrid systems of human and artificial agents (Floridi 2012; Crain 2018).

Additionally, due to the structure and operation of the data brokerage market, it is in many cases impossible to "trace any given datum to its original source" once it has been introduced to the marketplace (Crain 2018, 93). Reasons for this include trade secret protection; complex markets that "divorce" the data collection process from the selling and buying process; and the mix of large volumes of computationally generated information with "no 'real' empirical source" combined with genuine data (Crain 2018, 94).

The technical complexity and dynamism of ML algorithms make them prone to concerns of "agency laundering": a moral wrong which consists in distancing oneself from morally suspect actions, regardless of whether those actions were intended or not, by blaming the algorithm (Rubel et al. 2019). This is practiced by organisations as well as by individuals. Rubel et al. provide a straightforward and chilling example of agency laundering by Facebook:

> "Using Facebook's automated system, the ProPublica team found a user-generated category called "Jew hater" with over 2200 members. […] To help ProPublica find a larger audience (and hence have a better ad purchase), Facebook suggested a number of additional categories. […] ProPublica used the platform to select other profiles displaying anti-Semitic categories, and Facebook approved ProPublica's ad with minor changes. When ProPublica revealed the anti-Semitic categories and other news outlets reported similarly odious categories, Facebook responded by explaining that algorithms had created the categories based on user responses to target fields [and that] "[w]e never intended or anticipated this functionality being used this way" (Rubel et al. 2019, 1024–25).

Today, the failure to grasp the unintended effects of mass personal data processing and commercialisation, a familiar problem in the history of technology (Wiener 1950; Klee 1996; Benjamin 2019), is coupled with the limited explanations that most ML algorithms provide (Watson et al. 2019). This approach risks to favour avoidance of responsibility through "the computer said so" type of denial (Karppi 2018). This can lead field experts, such as clinicians, to avoid questioning the suggestion of an algorithm even when it may seem odd to them. The interplay between field experts and ML algorithms can prompt "epistemic vices" (Grote and Berens 2020), such as dogmatism or gullibility (Hauer 2019), and hinder the attribution of responsibility in distributed systems (Floridi 2016). To address this issue, Shah's analysis (2018) stresses that the risk that some stakeholders

may breach their responsibilities can be addressed, for example, by establishing separate bodies for the ethical oversight of algorithms (e.g. DeepMind Health established an Independent Review Panel with unfettered access to the company until Google halted it in 2019) (Murgia 2018). However, expecting a single oversight body, like a research ethics committee or institutional review board, to "be solely responsible for ensuring the rigour, utility, and probity of big data" is unrealistic (Lipworth et al. 2017, 8). Indeed, some have argued that these initiatives lack any sort of consistency and can rather lead to "ethics bluewashing", understood as.

> "implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions to appear more digitally ethical than one is." (Floridi 2019b, 187).

Faced with strict legal regimes, resourceful actors may also resort to so-called "ethics dumping" whereby unethical "processes, products or services" are exported to countries with weaker frameworks and enforcement mechanisms, after which the outcomes of such unethical activities are "imported back" (Floridi 2019b, 190).

There are a number of detailed approaches to establishing algorithmic accountability in the reviewed literature. While ML algorithms do require a level of technical intervention to improve their explainability, most approaches focus on normative interventions (Fink 2018). For example, Ananny and Crawford argue that, at least, providers of algorithms ought to facilitate public discourse about their technology (Ananny and Crawford 2018). Similarly, to address the issue of ad hoc ethical actions, some have claimed that accountability should first and foremost be addressed as a matter of convention (Dignum et al. 2018; Reddy et al. 2019).

Looking to fill the convention "gap", Buhmann et al. (2019) borrow from the seven principles for algorithms set out by the Association for Computing Machinery, claiming that through, *inter alia*, awareness of their algorithms, validation, and testing, an organisation should take responsibility for their algorithms regardless of how opaque they are (Malhotra et al. 2018). Decisions regarding the deployment of algorithms should incorporate factors such as desirability and the wider context in which they will operate, which should then lead to a more accountable "algorithmic culture" (Vedder and Naudts 2017, 219). To capture such considerations, "interactive and discursive fora and processes" with relevant stakeholders, as suggested by Buhmann et al., may prove a useful means (Buhmann et al. 2019, 13).

In the same vein, Binns (2018) focuses on the political-philosophical concept of "public reason". Considering that the processes for ascribing responsibility for the actions of an algorithm differ, both in nature and scope, in the public versus private sector, Binns calls for the establishment of a publicly shared framework (Binns 2018; see also Dignum et al. 2018), according to which algorithmic decisions should be able to withstand the same level of public scrutiny that human decision-making would receive. This approach has been echoed by many others in the reviewed literature (Ananny and Crawford 2018; Blacklaws 2018; Buhmann et al. 2019).

Problems relating to 'agency laundering' and 'ethics shirking' arise from the inadequacy of existing conceptual frameworks to trace and ascribe moral responsibility. As Floridi points out, when considering algorithmic systems and the impact of their actions.

> "we are dealing with DMAs [distributed moral actions] arising from morally neutral interactions of (potentially hybrid) networks of agents? In other words, who is responsible (*distributed moral responsibility*, DMR) for DMAs?", (Floridi 2016, 2).

Floridi's analysis suggests ascribing full moral responsibility "by default and overridably" to *all* the agents in the network which are causally relevant to the given action of the network. The proposed approach builds on the concepts of back-propagation from network theory, strict liability from jurisprudence, and common knowledge from epistemic logic. Notably, this approach decouples moral responsibility from the intentionality of the actors and from the very idea of punishment and reward for performing a given action, to focus instead on the need to rectify mistakes (back-propagation) and improve the ethical working of all the agents in the network.

9 Conclusion

This article builds on, and updates, previous research conducted by our group (Mittelstadt et al. 2016) to review relevant literature published since 2016 on the ethics of algorithms. Although that article is now inevitably outdated in terms of specific references and detailed information about the literature reviewed, the map, and the six categories that it provides, have withstood the test of time and remain a valuable tool to scope ethics of algorithms as an area of research, with a growing body of literature focusing on each of the six categories contributing either to refine our understanding of existing problems or to provide solutions to address them.

Since 2016, the ethics of algorithms has become a central topic of discussion among scholars, technology providers, and policymakers. The debate has gained traction also because of the so-called "summer of AI", and with it the pervasive use of ML algorithms. Many of the ethical questions analysed in this article and the literature it reviews have been addressed in national and international ethical guidelines and principles, like the aforementioned European

Commission's European Group on Ethics in Science and Technologies, the UK's House of Lords Artificial Intelligence Committee (Floridi and Cowls 2019), and the OECD principles on AI (OECD 2019).

One aspect that was not explicitly captured by the original map, and which is becoming a central point of discussion in the relevant literature, is the increasing focus on the use of algorithms, AI and digital technologies more broadly, to deliver socially good outcomes (Hager et al. 2019) (Floridi et al. 2020; Cowls et al. 2021). While it is true, at least in principle, that any initiative aimed at using algorithms for social good should address satisfactorily the risks that each of the six categories in the map identifies, there is also a growing debate on the principles and criteria that should inform the design and governance of algorithms, and digital technologies more broadly, for the explicit purpose of social good.

Ethical analyses are necessary to mitigate the risks while harnessing the potential for good of these technologies, insofar as they serve the twin goals of clarifying the nature of the ethical risks and of the potential for good of algorithms and digital technologies, and translating (Taddeo and Floridi 2018b; Morley et al. 2019a, b) this understanding into sound, actionable guidance for the governance of the design and use of digital artefacts.

Appendix

Methodology

Four databases of academic literature were systematically queried (see: Table 1) to identify literature discussing ethics and algorithms. Four keywords were used to describe an algorithm: 'algorithm', 'machine learning', 'software' and 'computer program'.[3] The search was limited to publications from November 2016 to March 2020.

The search identified 4891 unique papers for review.[4] After an initial review of title/abstract, 180 papers were selected for a full review. Of these, 62 were rejected as off-topic, leaving 118 articles for full review.

Another 37 articles and books were reviewed and referenced in this paper to provide additional information regarding specific ethical issues and solutions (e.g., technical details, examples and tools). These were sourced from the bibliographies of the 118 articles we reviewed as well as provided on an ad-hoc basis when agreed upon by the authors as being helpful for clarification.

[3] The literature search was limited to English language articles in peer-reviewed journals and conference proceedings.

[4] Many of which were purely technical in nature, especially for "discrimination" and "(transparency OR scrutability OR opacity)".

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp 308–18. Vienna Austria: ACM. https://doi.org/10.1145/2976749.2978318. Accessed 24 Aug 2020

Abebe R, Barocas S, Kleinberg J, Levy K, Raghavan M, Robinson DG (2020) Roles for computing in social change. https://arxiv.org/pdf/1912.04883.pdf. Accessed 24 Aug 2020

Aggarwal N (2020) The norms of algorithmic credit scoring. SSRN Electron J. https://doi.org/10.2139/ssrn.3569083

Allen A (2011) Unpopular privacy what must we hide? Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195141375.001.0001

Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20(3):973–989. https://doi.org/10.1177/1461444816676645

Angwin J, Larson J, Mattu S, Lauren K (2016) Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 24 Aug 2020

Arnold M, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilovic A, Nair R et al (2019) FactSheets: increasing trust in AI services through supplier's declarations of conformity. ArXiv:1808.07261. http://arxiv.org/abs/1808.07261. Accessed 24 Aug 2020

Bambauer J, Zarsky T (2018) The algorithmic game. Notre Dame Law Rev 94(1):1–47

Barocas S, Selbst AD (2016) Big data's disparate impact. SSRN Electron J. https://doi.org/10.2139/ssrn.2477899

Bauer WA, Dubljević V (2020) AI assistants and the paradox of internal automaticity. Neuroethics 13(3):303–310. https://doi.org/10.1007/s12152-019-09423-6

Baumer EPS (2017) Toward human-centered algorithm design. Big Data Soc 4(2):205395171771885

Beer D (2017) The social power of algorithms. Inform Commun Soc 20(1):1–13. https://doi.org/10.1080/1369118X.2016.1216147

Benjamin R (2019) Race after technology: abolitionist tools for the new jim code. Polity, Medford

Benjamin R (2020) 2020 Vision: reimagining the default settings of technology and society. https://iclr.cc/virtual_2020/speaker_3.html. Accessed 24 Aug 2020

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res. https://doi.org/10.1177/0049124118782533

Binns R (2018) Fairness in machine learning: lessons from political philosophy. ArXiv:1712.03586. http://arxiv.org/abs/1712.03586. Accessed 24 Aug 2020

Blacklaws C (2018) Algorithms: transparency and accountability. Philos Trans R Soc A Math Phys Eng Sci 376(2128):20170351. https://doi.org/10.1098/rsta.2017.0351

Blyth CR (1972) On Simpson's paradox and the sure-thing principle. J Am Stat Assoc 67(338):364–366. https://doi.org/10.1080/01621459.1972.10482387

Boyd D, Crawford K (2012) Critical questions for big data. Inform Commun Soc 15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878

Buhmann A, Paßmann J, Fieseler C (2019) Managing algorithmic accountability: balancing reputational concerns, engagement strategies, and the potential of rational discourse. J Bus Ethics. https://doi.org/10.1007/s10551-019-04226-4

Burke R (2017) Multisided fairness for recommendation. ArXiv:1707.00093. http://arxiv.org/abs/1707.00093. Accessed 24 Aug 2020

Burrell J (2016) How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc 3(1):205395171562251. https://doi.org/10.1177/2053951715622512

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H et al. (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. ArXiv:2004.07213 [Cs]. http://arxiv.org/abs/2004.07213. Accessed 24 Aug 2020

Chakraborty A, Patro GK, Ganguly N, Gummadi KP, Loiseau P (2019) Equality of voice: towards fair representation in crowdsourced top-K recommendations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, 129–38. Atlanta, GA, USA: ACM Press. https://doi.org/10.1145/3287560.3287570

Cohen J (2000) Examined lives: informational privacy and the subject as object. Georgetown Law Faculty Publications and Other Works, January. https://scholarship.law.georgetown.edu/facpub/810. Accessed 24 Aug 2020

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. ArXiv:1808.00023. http://arxiv.org/abs/1808.00023. Accessed 24 Aug 2020

Cowls J, Tsamados A, Taddeo M, Floridi L (2021) A definition, benchmark and database of AI for social good initiatives. Nat Mach Intell

Crain M (2018) The limits of transparency: data brokers and commodification. New Media Soc 20(1):88–104. https://doi.org/10.1177/1461444816657096

Cummings M (2012) Automation bias in intelligent time critical decision support systems. In: AIAA 1st Intelligent Systems Technical Conference. Chicago, Illinois: American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

Dahl ES (2018) Appraising Black-boxed technology: the positive prospects. Philos Technol 31(4):571–591. https://doi.org/10.1007/s13347-017-0275-1

Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 4691–97. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/654

Datta A, Tschantz MC, Datta A (2015) Automated experiments on Ad privacy settings. Proc Priv Enhanc Technol 2015(1):92–112. https://doi.org/10.1515/popets-2015-0007

Davis E, Marcus G (2019) Rebooting AI: building artificial intelligence we can trust. Pantheon Books, New York

Diakopoulos N, Koliska M (2017) Algorithmic transparency in the news media. Digit Journal 5(7):809–828. https://doi.org/10.1080/21670811.2016.1208053

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. ArXiv:1702.08608. http://arxiv.org/abs/1702.08608. Accessed 24 Aug 2020

Edwards L, Veale M (2017) Slave to the algorithm? Why a right to explanationn is probably not the remedy you are looking for. SSRN Electron J. https://doi.org/10.2139/ssrn.2972855

Ekstrand M, Levy K (2018) FAT* Network. https://fatconference.org/network. Accessed 24 August 2020

Eubanks V (2017) Automating inequality: how high-tech tools profile, police, and punish the poor, 1st edn. St. Martin's Press, New York

Fink K (2018) Opening the government's black boxes: freedom of information and algorithmic accountability. Inform Commun Soc 21(10):1453–1471. https://doi.org/10.1080/1369118X.2017.1330418

Floridi L (2012) Distributed morality in an information society. Sci Eng Ethics 19(3):727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi L (2016) Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philos Trans R Soc A Math Phys Eng Sci 374(2083):20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi L (2017) Infraethics–on the conditions of possibility of morality. Philos Technol 30(4):391–394. https://doi.org/10.1007/s13347-017-0291-1

Floridi L (2019b) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi L (2019a) What the near future of artificial intelligence could be. Philos Technol 32(1):1–15. https://doi.org/10.1007/s13347-019-00345-y

Floridi L, Cowls J (2019) A unified framework of five principles for ai in society. Harvard Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1

Floridi L, Taddeo M (2016) What is data ethics? Philos Trans R Soc A: Math Phys Eng Sci 374(2083):20160360. https://doi.org/10.1098/rsta.2016.0360

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi L, Cowls J, King TC, Taddeo M (2020) How to design AI for social good: seven essential factors. Sci Eng Ethics 26(3):1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2017) Predictably unequal? The effects of machine learning on credit markets. SSRN Electron J. https://doi.org/10.2139/ssrn.3072038

Gajane P, Pechenizkiy M (2018) On formalizing fairness in prediction with machine learning. ArXiv:1710.03184. http://arxiv.org/abs/1710.03184. Accessed 24 Aug 2020

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2020) Datasheets for datasets. ArXiv:1803.09010. http://arxiv.org/abs/1803.09010. Accessed 1 Aug 2020

Gillis TB, Spiess J (2019) Big data and discrimination. Univ Chicago Law Rev 459

Grant MJ, Booth A (2009) Types and associated methodologies: a typology of reviews. Health Inform Lib J 26(2):91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Green B, Chen Y (2019) Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT*'19, 90–99. Atlanta, GA, USA: ACM Press. https://doi.org/10.1145/3287560.3287563

Green B, Viljoen S (2020) Algorithmic realism: expanding the boundaries of algorithmic thought. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 19–31. Barcelona Spain: ACM. https://doi.org/10.1145/3351095.3372840

Grgić-Hlača N, Redmiles EM, Gummadi KP, Weller A (2018) Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. ArXiv:1802.09548. http://arxiv.org/abs/1802.09548. Accessed 24 Aug 2020

Grote T, Berens P (2020) On the ethics of algorithmic decision-making in healthcare. J Med Ethics 46(3):205–211. https://doi.org/10.1136/medethics-2019-105586

Hager GD, Drobnis A, Fang F, Ghani R, Greenwald A, Lyons T, Parkes DC et al (2019) Artificial intelligence for social good. ArXiv:1901.05406 http://arxiv.org/abs/1901.05406. Accessed 24 Aug 2020

Harwell D (2020) Dating apps need women. advertisers need diversity. AI companies offer a solution: fake people. Washington Post

Hauer T (2019) Society caught in a labyrinth of algorithms: disputes, promises, and limitations of the new order of things. Society 56(3):222–230. https://doi.org/10.1007/s12115-019-00358-5

Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R, Pineau J (2018) Ethical challenges in data-driven dialogue systems. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 123–29. New Orleans LA USA: ACM. https://doi.org/10.1145/3278721.3278777

Hill RK (2016) What an algorithm is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5

Hleg AI (2019) Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 24 Aug 2020

Hoffmann AL, Roberts ST, Wolf CT, Wood S (2018) Beyond fairness, accountability, and transparency in the ethics of algorithms: contributions and perspectives from LIS. Proc Assoc Inform Sci Technol 55(1):694–696. https://doi.org/10.1002/pra2.2018.14505501084

Hu M (2017) Algorithmic Jim Crow. Fordham Law Review. https://ir.lawnet.fordham.edu/flr/vol86/iss2/13/. Accessed 24 Aug 2020

Hutson M (2019) Bringing machine learning to the masses. Science 365(6452):416–417. https://doi.org/10.1126/science.365.6452.416

ICO (2020) ICO and The Turing Consultation on Explaining AI Decisions Guidance. ICO. 30 March 2020. https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/. Accessed 24 Aug 2020

James G, Witten G, Hastie T, Tibshirani R (2013) An Introduction to statistical learning. Springer, New York

Karppi T (2018) The computer said so: on the ethics, effectiveness, and cultural techniques of predictive policing. Soc Media Soc 4(2):205630511876829. https://doi.org/10.1177/2056305118768296

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. ArXiv:1812.04948. http://arxiv.org/abs/1812.04948. Accessed 24 Aug 2020

Katell M, Young M, Dailey D, Herman B, Guetler V, Tam A, Binz C, Raz D, Krafft PM (2020) Toward Situated interventions for algorithmic equity: lessons from the field. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 45–55. Barcelona Spain: ACM. https://doi.org/10.1145/3351095.3372874

King G, Persily N (2020) Unprecedented Facebook URLs dataset now available for academic research through social science one. 2020. Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One

Kizilcec R (2016) How much information? In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp 2390–2395. https://doi.org/10.1145/2858036.2858402

Klee R (1996) Introduction to the philosophy of science: cutting nature at its seams. Oxford University Press, Oxford

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. ArXiv:1609.05807. http://arxiv.org/abs/1609.05807. Accessed 24 Aug 2020

Kortylewski A, Egger B, Schneider A, Gerig T, Morel-Forster F, Vetter T (2019) Analyzing and Reducing the damage of dataset bias to face recognition with synthetic data. http://openaccess.thecvf.com/content_CVPRW_2019/html/BEFA/Kortylewski_Analyzing_and_Reducing_the_Damage_of_Dataset_Bias_to_Face_CVPRW_2019_paper.html. Accessed 24 Aug 2020

Labati RD, Genovese A, Muñoz E, Piuri V, Scotti F, Sforza G (2016) Biometric recognition in automated border control: a survey. ACM Comput Surv 49(2):1–39. https://doi.org/10.1145/2933241

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag Sci 65(7):2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Larson B (2017) Gender as a variable in natural-language processing: ethical considerations. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 1–11. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1601

Lee MK (2018) Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc 5(1):205395171875668. https://doi.org/10.1177/2053951718756684

Lee TN (2018) Detecting racial bias in algorithms and machine learning. J Inform Commun Ethics Soc 16(3):252–260. https://doi.org/10.1108/JICES-06-2018-0056

Lee MS, Floridi L (2020) Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. SSRN Electron J. https://doi.org/10.2139/ssrn.3559407

Lee MK, Kim JT, Lizarondo L (2017) A human-centered approach to algorithmic services: considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems—CHI '17, 3365–76. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025884

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. Philos Technol 31(4):611–627. https://doi.org/10.1007/s13347-017-0279-x

Lewis D (2019) Social Credit case study: city citizen scores in Xiamen and Fuzhou'. Medium: Berkman Klein Center Collection. 8 October 2019. https://medium.com/berkman-klein-center/social-credit-case-study-city-citizen-scores-in-xiamen-and-fuzhou-2a65feb2bbb3. Accessed 10 Oct 2020

Lipworth W, Mason PH, Kerridge I, Ioannidis JPA (2017) Ethics and epistemology in big data research. J Bioethical Inq 14(4):489–500. https://doi.org/10.1007/s11673-017-9771-3

Magalhães JC (2018) Do algorithms shape character? Considering algorithmic ethical subjectivation. Soc Media Soc 4(2):205630511876830. https://doi.org/10.1177/2056305118768301

Malhotra C, Kotwal V, Dalal S (2018) Ethical framework for machine learning. In: 2018 ITU Kaleidoscope: machine

learning for a 5G Future (ITU K), 1–8. Santa Fe: IEEE. https://doi.org/10.23919/ITU-WT.2018.8597767
Martin K (2019) Ethical implications and accountability of algorithms. J Bus Ethics 160(4):835–850. https://doi.org/10.1007/s10551-018-3921-3
Mayson SG (2019) 'Bias In, Bias Out'. *Yale Law Journal*, no. 128. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004. Accessed 24 Aug 2020
Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. AI Soc. https://doi.org/10.1007/s00146-020-00950-y
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679
Mojsilovic A (2018) Introducing AI explainability 360. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/. Accessed 24 Aug 2020
Möller J, Trilling D, Helberger N, van Es B (2018) Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. Inform Commun Soc 21(7):959–977. https://doi.org/10.1080/1369118X.2018.1444076
Morley J, Floridi L, Kinsey L, Elhalal A (2019) From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Sci Eng Ethics. https://doi.org/10.1007/s11948-019-00165-5
Morley J, Machado C, Burr C, Cowls J, Taddeo M, Floridi L (2019) The debate on the ethics of ai in health care: a reconstruction and critical review. SSRN Electron J. https://doi.org/10.2139/ssrn.3486518
Murgia M (2018) DeepMind's move to transfer health unit to Google Stirs data fears. Financial Times, New York, p 2018
Noble SU (2018) Algorithms of oppression: how search engines reinforce racism. New York University Press, New York
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342
Ochigame R (2019) The invention of "Ethical AI", 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/. Accessed 24 Aug 2020
OECD (2019) Recommendation of the council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed 24 Aug 2020
Olhede SC, Wolfe PJ (2018) The growing ubiquity of algorithms in society: implications, impacts and innovations. Philos Trans R Soc A Math Phys Eng Sci 376(2128):20170364. https://doi.org/10.1098/rsta.2017.0364
Olteanu A, Castillo C, Diaz F, Kiciman E (2016) Social data: biases, methodological pitfalls, and ethical boundaries. SSRN Electron J. https://doi.org/10.2139/ssrn.2886526
Oswald M (2018) Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. Philos Trans R Soc A: Math Phys Eng Sci 376(2128):20170359. https://doi.org/10.1098/rsta.2017.0359
Paraschakis D (2017) Towards an ethical recommendation framework. In: 2017 11th International Conference on Research Challenges in Information Science (RCIS), 211–20. Brighton, United Kingdom: IEEE. https://doi.org/10.1109/RCIS.2017.7956539
Paraschakis D (2018) Algorithmic and ethical aspects of recommender systems in E-commerce. Malmö Universitet, Malmö
Perra N, Rocha LEC (2019) Modelling opinion dynamics in the age of algorithmic personalisation. Sci Rep 9(1):7261. https://doi.org/10.1038/s41598-019-43830-2
Perrault R, Yoav S, Brynjolfsson E, Jack C, Etchmendy J, Grosz B, Terah L, James M, Saurabh M, Carlos NJ (2019) Artificial Intelligence Index Report 2019
Prates MOR, Avelar PH, Lamb LC (2019) Assessing gender bias in machine translation: a case study with google translate. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04144-6
Dignum V, Lopez-Sanchez M, Micalizio R, Pavón J, Slavkovik M, Smakman M, van Steenbergen M et al (2018) Ethics by design: necessity or curse? In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society—AIES '18, 60–66. New Orleans, LA, USA: ACM Press. https://doi.org/10.1145/3278721.3278745
Rachels J (1975) Why privacy is important. Philos Public Aff 4(4):323–333
Rahwan I (2018) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14. https://doi.org/10.1007/s10676-017-9430-8
Ras G, van Gerven M, Haselager P (2018) Explanation methods in deep learning: users, values, concerns and challenges. ArXiv:1803.07517. http://arxiv.org/abs/1803.07517. Accessed 24 Aug 2020
Reddy E, Cakici B, Ballestero A (2019) Beyond mystery: putting algorithmic accountability in context. Big Data Soc 6(1):205395171982685. https://doi.org/10.1177/2053951719826856
Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability'. AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf. Accessed 24 Aug 2020
Richardson R, Schultz J, Crawford K (2019) Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423. Accessed 24 Aug 2020
Robbins S (2019) A misdirected principle with a catch: explicability for AI. Mind Mach 29(4):495–514. https://doi.org/10.1007/s11023-019-09509-3
Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L (2019) The Chinese approach to artificial intelligence: an analysis of policy and regulation. SSRN Electron J. https://doi.org/10.2139/ssrn.3469784
Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L (2020) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI Soc. https://doi.org/10.1007/s00146-020-00992-2
Rössler B (2015) The value of privacy. https://philpapers.org/rec/ROSTVO-9. Accessed 24 Aug 2020
Rubel A, Castro C, Pham A (2019) Agency laundering and information technologies. Ethical Theory Moral Pract 22(4):1017–1041. https://doi.org/10.1007/s10677-019-10030-w
Sandvig C, Hamilton K, Karahalios K, Langbort C (2016) When the algorithm itself is a racist: diagnosing ethical harm in the basic components of software. Int J Commun 10:4972–4990
Saxena N, Huang K, DeFilippis E, Radanovic G, Parkes D, Liu Y (2019) How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. ArXiv:1811.03654. http://arxiv.org/abs/1811.03654. Accessed 24 Aug 2020
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, 59–68. Atlanta, GA, USA: ACM Press. https://doi.org/10.1145/3287560.3287598
Shah H (2018) Algorithmic accountability. Philos Trans R Soc A: Math Phys Eng Sci 376(2128):20170362. https://doi.org/10.1098/rsta.2017.0362

Shin D, Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. Comput Hum Behav 98(September):277–284. https://doi.org/10.1016/j.chb.2019.04.019

Sloan RH, Warner R (2018) When is an algorithm transparent? predictive analytics, privacy, and public policy. IEEE Secur Priv 16(3):18–25. https://doi.org/10.1109/MSP.2018.2701166

Stilgoe J (2018) Machine learning, social learning and the governance of self-driving cars. Soc Stud Sci 48(1):25–56. https://doi.org/10.1177/0306312717741687

Szegedy C, Wojciech Z, Ilya S, Joan B, Dumitru E, Ian G, Rob F (2014) Intriguing Properties of Neural Networks. ArXiv:1312.6199 [Cs]. http://arxiv.org/abs/1312.6199. Accessed 18 July 2020

Taddeo M, Floridi L (2018a) Regulate artificial intelligence to avert cyber arms race. Nature 556(7701):296–298. https://doi.org/10.1038/d41586-018-04602-6

Taddeo M, Floridi L (2018b) How AI can be a force for good. Science 361(6404):751–752. https://doi.org/10.1126/science.aat5991

Taddeo M, McCutcheon T, Floridi L (2019) Trusting artificial intelligence in cybersecurity is a double-edged sword. Nat Mach Intell 1(12):557–560. https://doi.org/10.1038/s42256-019-0109-1

Taylor L, Floridi L, van der Sloot B (eds) (2017) Group privacy: new challenges of data technologies. Springer, Berlin Heidelberg, New York

Tickle AB, Andrews R, Golea M, Diederich J (1998) The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Trans Neural Netw 9(6):1057–1068. https://doi.org/10.1109/72.728352

Turilli M, Floridi L (2009) The ethics of information transparency. Ethics Inf Technol 11(2):105–112. https://doi.org/10.1007/s10676-009-9187-9

Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142. https://doi.org/10.1145/1968.1972

Veale M, Binns R (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2):205395171774353. https://doi.org/10.1177/2053951717743530

Vedder A, Naudts L (2017) Accountability for the use of algorithms in a big data environment. Int Rev Law Comput Technol 31(2):206–224. https://doi.org/10.1080/13600869.2017.1298547

Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, Dow M, Ohno-Machado L (2017) Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the united states: genome privacy in biomedical research. Ann N Y Acad Sci 1387(1):73–83. https://doi.org/10.1111/nyas.13259

Watson D, Floridi L (2020) The explanation game: a formal framework for interpretable machine learning. SSRN Electron J. https://doi.org/10.2139/ssrn.3509737

Watson DS, Krutzinna J, Bruce IN, Griffiths CEM, McInnes IB, Barnes MR, Floridi L (2019) Clinical applications of machine learning algorithms: beyond the black box. BMJ. https://doi.org/10.1136/bmj.l886

Webb H, Patel M, Rovatsos M, Davoust A, Ceppi S, Koene A, Dowthwaite L, Portillo V, Jirotka M, Cano M (2019) "It would be pretty immoral to choose a random algorithm": opening up algorithmic interpretability and transparency. J Inform Commun Ethics Soc 17(2):210–228. https://doi.org/10.1108/JICES-11-2018-0092

Weller A (2019) Transparency: motivations and challenges. ArXiv:1708.01870. http://arxiv.org/abs/1708.01870. Accessed 24 Aug 2020

Wexler J (2018) The what-if tool: code-free probing of machine learning models. https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html. Accessed 24 Aug 2020

Whitman M, Hsiang C-y, Roark K (2018) Potential for participatory big data ethics and algorithm design: a scoping mapping review. In: Proceedings of the 15th Participatory Design Conference on Short Papers, Situated Actions, Workshops and Tutorial - PDC '18, 1–6. Hasselt and Genk, Belgium: ACM Press. https://doi.org/10.1145/3210604.3210644

Wiener N (1950) The human use of human beings.

Winner L (1980) Do artifacts have politics? Modern Techn Probl Oppor 109(1):121–136

Wong P-H (2019) Democratizing algorithmic fairness. Philos Technol. https://doi.org/10.1007/s13347-019-00355-w

Xian Z, Li Q, Huang X, Li L (2017) New SVD-based collaborative filtering algorithms with differential privacy. J Intell Fuzzy Syst 33(4):2133–2144. https://doi.org/10.3233/JIFS-162053

Xu D, Yuan S, Zhang L, Wu X (2018) FairGAN: fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), 570–75. Seattle, WA, USA: IEEE. https://doi.org/10.1109/BigData.2018.8622525

Yang G-Z, Bellingham J, Dupont PE, Fischer P, Floridi L, Full R, Jacobstein R et al (2018) The grand challenges of science robotics. Sci Robot 3(14):eaar7650. https://doi.org/10.1126/scirobotics.aar7650

Yampolskiy RV (2018) Artificial intelligence safety and security

Yu M, Du G (2019) Why are Chinese courts turning to AI?' The Diplomat. 19 January 2019. https://thediplomat.com/2019/01/why-are-chinese-courts-turning-to-ai/. Accessed 24 Aug 2020

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32(4):661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhou Na, Zhang C-T, Lv H-Y, Hao C-X, Li T-J, Zhu J-J, Zhu H et al (2019) Concordance study between IBM Watson for oncology and clinical practice for patients with cancer in China. Oncologist 24(6):812–819. https://doi.org/10.1634/theoncologist.2018-0255