Commentary

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–8
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2053951720983865
journals.sagepub.com/home/bds
SAGE

The algorithm audit: Scoring the algorithms that score us

Shea Brown[1,2], Jovana Davidovic[3,2] and Ali Hasan[3,2]

Abstract

In recent years, the ethical impact of AI has been increasingly scrutinized, with public scandals emerging over biased outcomes, lack of transparency, and the misuse of data. This has led to a growing mistrust of AI and increased calls for mandated ethical audits of algorithms. Current proposals for ethical assessment of algorithms are either too high level to be put into practice without further guidance, or they focus on very specific and technical notions of fairness or transparency that do not consider multiple stakeholders or the broader social context. In this article, we present an auditing framework to guide the ethical assessment of an algorithm. The audit instrument itself is comprised of three elements: a list of possible interests of stakeholders affected by the algorithm, an assessment of metrics that describe key ethically salient features of the algorithm, and a relevancy matrix that connects the assessed metrics to stakeholder interests. The proposed audit instrument yields an ethical evaluation of an algorithm that could be used by regulators and others interested in doing due diligence, while paying careful attention to the complex societal context within which the algorithm is deployed.

Keywords

Algorithm audits, algorithm ethics, ethics, ethics of AI, machine learning, machine learning and ethics

Introduction

The rapid development of artificial intelligence (AI) and machine learning has led to powerful algorithms that have the potential to improve lives on an unprecedented scale. But, with greater capabilities comes greater potential for harm. Algorithms, and especially machine learning algorithms, are more and more often used to supplant or augment human decision-making in a way that affects human lives, interests, opportunities, and rights. Algorithms assess job applicants, loan applicants, bail applicants, student tests, fitness for rental properties, etc. (Eubanks, 2018). In recent years, the ethical impact of AI has been increasingly scrutinized, with public scandals emerging over the lack of transparency, misuse of data, and the propagation of systemic racism (e.g., Benjamin, 2019; Noble, 2018; O'Neil, 2016; Prabhu and Birhane, 2020; Whittaker et al., 2018).

In response to these growing concerns, nearly every research organization that deals with the ethics of AI has called for ethical auditing of algorithms. A recent example is the EU High-Level Expert Group on Artificial Intelligence, who emphasized this need in their draft, "Ethics Guidelines for Trustworthy AI". In the US, "automated decision system impact assessments" have been proposed by Congress as part of the Algorithmic Accountability Act of 2019.

While it is clear why audits could help build trust with the public, how these "algorithm audits" are to be done is still an open question and an area of active research. Current proposals are either too high level to be put into practice without further guidance (Barocas et al., 2013; Floridi et al., 2018; Mittelstadt et al., 2016; Raji et al., 2020; Sandvig et al., 2014), or they focus on narrow and often technical notions of fairness, bias, or transparency (Mitchell et al., 2019) and do not consider multiple stakeholders or the

[1]Department of Physics & Astronomy, University of Iowa, Iowa City, IA, USA
[2]BABL AI Research, Iowa City, IA, USA
[3]Department of Philosophy, University of Iowa, Iowa City, IA, USA

Corresponding author:
Jovana Davidovic, University of Iowa, 273 EPB, Iowa City, IA 52245, USA.
Email: jovana-davidovic@uiowa.edu

broader social context (Mittelstadt, 2019; Selbst et al., 2018). Third-party ethical assessments of algorithms are clearly sorely needed, but developing a mechanism for such assessments is a complex task (Ada Lovelace Institute, 2020; Brundage et al., 2020).

In this article, we present a sketch of a framework for an algorithm audit, with an understanding that what we propose here requires further development and discussion in order to be functionally complete and ready to implement. We start by defining "ethical audits". Next, we describe the preliminary steps of the audit: identifying the goals of the audit, and describing the context of the algorithm. The context of the algorithm is one of the most overlooked elements of ethical audit proposals to date. We then describe our audit instrument, which depends on three building blocks: a list of possible interests of stakeholders affected by the algorithm, an assessment of the metrics that describe key ethically salient features of the algorithm, and a relevancy matrix that connects the metrics to stakeholder interests. Finally, we describe what an outcome of an algorithm audit would look like, yielding an evaluation of an algorithm that could be used by regulators and others interested in doing their due diligence.

The aim of this article is to propose one way to operationalize high-level ethical analyses of algorithms by suggesting an auditing instrument which translates those ethical analyses into practical steps. There are many other key questions about what auditing mechanisms ought to look like, especially when they are used or intended to be used by regulators. These include structural questions about power dynamics, questions about how regulatory agencies performing audits will be organized, and who will perform the audits (Ada Lovelace Institute, 2020; Brundage et al., 2020; Schiff et al., 2020). We put aside those questions for now, since the auditing instrument we propose here is not solely designed for purposes of regulation by the state.

What is an algorithm audit?

As they are typically discussed in the literature, audits involve collecting data about the behavior of an algorithm as it is used in a particular context, and then using that data to assess whether the behavior is negatively impacting some interests (or rights) of people affected by that algorithm. In the case of algorithms that assign some sort of score to humans, such as risk scores (Obermeyer et al., 2019) or credit scores (Deville, 2019), audits have focused on issues of unfair treatment of certain groups based on potential bias. For algorithms that track online behavior in order to personalize (limit) ads and products, audits have focused primarily on issues of transparency or autonomy. Audits of some algorithms, such as facial recognition or affect recognition (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019), have focused not only on bias, but also on the potential for abuse. The audit instrument we sketch here is meant to be more comprehensive and broadly applicable.

In line with literature, we define ethical algorithm audits as *assessments of the algorithm's negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts*. We focus on the negative impacts because those are the impacts that regulators are primarily interested in identifying and limiting, and because those impacts are more immediately related to risk management.

Preliminary analysis

The key preliminary steps for our proposed audit mechanism include identifying the way in which the audit will be used, and describing and circumscribing the context of an algorithm for the purposes of the audit.

(i) *Audit purpose*: An ethical audit might be used in at least three general ways. First, it could be used by regulators to assess whether some algorithm meets legal standards or internal policies. For example, regulators might be interested in whether a bank's lending algorithm meets Federal Housing Administration standards. Second, an algorithm audit might also be used by algorithm vendors and buyers to mitigate or control ethical and reputational risks and to identify ways to remedy those risks. Finally, stakeholders might be interested in a general ethical assessment of an algorithm so as to make informed choices about voting, investing, engaging with certain companies, etc. The audit framework we develop can produce a range of assessment detail and it is meant for a variety of uses that broadly fall under these three categories (regulatory, risk management, general ethical assessment).

(ii) *Context*: Understanding the context within which the algorithm is deployed means assessing and understanding a range of broader social and political facts about its stated purpose. The context of an algorithm is the socio-technical setting within which it is deployed. It might include the process of development of the algorithm, the process of preparing the data for the training algorithm, the process of delivering an algorithm to its primary user, and often, most importantly, the setting within which it is used. The analysis of the context of the algorithm might also include the dynamics of the algorithm's commercial trade, including whether it is open sourced or licensed. It is these

> contextual aspects of the algorithm's development and delivery that often account for much of the negative impacts of the algorithm. The ethical audits proposed to date often overlook the relevance of the social context for the purposes of ethical evaluation of the algorithm. To illustrate, the negative impacts of a loan risk tool do not simply depend on whether the algorithm is statistically biased against, for example, some minority group because of biased training data; possibly more importantly, the harm emerges from the way a loan officer decides to use that loan risk tool in her decision whether to give out a loan to an applicant.

The context of the algorithm is important for (a) identifying features of the algorithm to focus on (what we call "metrics"), (b) evaluating how well an algorithm does with respect to those features, and (c) identifying a complete list of the relevant stakeholders and their interests. The metrics and stakeholder interests are in turn the key building blocks of our proposed audit tool and as such clearly describing the context is an essential preliminary step in performing the audit. Technically identical algorithmic structures can yield vastly different audit outcomes depending on how the algorithm is used and/or developed.

The key questions to ask when describing the context of the algorithm include (Gebru et al., 2018; Mitchell et al., 2019): What is the stated purpose of the algorithm? Who is deploying the algorithm and what is their understanding of the purpose? How are the outputs of the algorithm used? Are these actions fully automated, or is there a human-in-the-loop and how does the human intervene or facilitate the loop? Is the algorithm static, or is it updated? If updated, at what cadence? How well is the group the algorithm is applied to represented in the training data? What common socio-political harms underpin and result from human decision-making that the algorithm is augmenting or supplanting? Answers to these and similar questions form a picture of the context within which the algorithm is deployed and thus the background against which the algorithm is to be evaluated.

Basic elements of the audit tool

With a clear picture of audit's purpose and context, we can deploy the audit instrument. The basic building blocks that are needed to perform the algorithm assessment include (1) a comprehensive list of relevant stakeholder interests and (2) a measure of the algorithm's potentially ethically relevant attributes (metrics). A clear description of the context is needed both to generate a list of stakeholder interests (1) and to *evaluate* the key features of the algorithm, i.e. metrics (2). Once steps (1) and (2) are completed we can (3) evaluate the relevance of a good or bad performance of an algorithm on some metric for each stakeholder interest. We can then use the metrics score (2) and the relevancy score (3) to determine the impact of the algorithm on stakeholder interests. We describe this process in more detail below.

Stakeholder interests & rights

An obvious major component that must be considered in any ethical audit is the rights and interests of the stakeholders. A stakeholder is anyone who could be affected by the use of an algorithm in some context. For example, consider an AI algorithm used to automatically assess student writing in an English class. In this case, stakeholders would include the students, parents, teachers, the school or university administering the course, the vendor that created the algorithm, and any regulatory body operating in this domain (e.g., U.S. Department of Education). One stakeholder that is persistent throughout almost all examples is society at large, and the interests of society can be captured through a variety of mechanisms (e.g. see the list of collective and societal harms from the Future of Privacy Forum, 2017). Given the context of a specific algorithm, it is apparent that not all of the interests of every stakeholder are equally affected by a flaw in some aspect of the algorithm (metric). A simple example is the right to privacy, which might be highly affected by a high potential for abuse in facial recognition systems, but less so in automated scoring algorithms used for grading students. Another example could be autonomy, which is threatened more by online filters (like which job advertisements you are served by Google) than by, e.g., healthcare diagnostic algorithms. To capture this, we propose a relevancy matrix between metrics and stakeholder interests at a later step. For now, all that matters is that we enumerate all relevant stakeholders and their interests that might even just plausibly be affected by the use of some algorithm. It is also very important to note that different stakeholders in the same category (e.g. students, loan applicants, those up for parole, digital passport users) are often affected in very different ways by the same algorithm and often on the basis of race, ethnicity, gender, religion, or sexual orientation (Benjamin, 2019). This is exactly why we argue that understanding the context of the algorithm is a precursor to being able to not only enumerate stakeholder interests generally—based on the kind of engagement a particular subject might have with a particular algorithm—but also to be able to identify particular sub-categories of stakeholders whose identification is relevant for ethical assessment of an algorithm (e.g. students of color, Hispanic loan applicants, male African-Americans up for parole, digital passport users who wear a hijab). These stakeholders might face particular threats, and context allows us to be particularly cognizant of not thinking that groups of stakeholders are homogeneous entities that will be negatively or positively affected simply in virtue of the type of engagement with an algorithm (e.g. student and grading algorithm), but also based on socio-political and socio-technical facts and power dynamics (Benjamin, 2019; D'Ignazio and Klein, 2020; Mohamed et al., 2020).

Metrics

After enumerating the interests that could possibly come under threat, the audit process turns to the algorithm itself. By an algorithm we do not mean simply the mathematical operations that make up the input–output function, but the larger socio-technical system that surrounds this function (Selbst et al., 2018). This includes the nature of the inputs, such as how they are gathered and fed into the function, the technical nature of the function itself, such as the model architecture and its performance and stability over time, and all other engineering details that govern its behavior. Importantly, a complete description of the algorithm also includes facts about how the output of the function is used in decision-making, and whether the actions taken are done so autonomously or with a human-in-the-loop. If actions are taken autonomously based on some threshold value of the output, details of how this threshold was decided upon and justified are important to the auditing process as well. Some of this information can be extracted directly from the context, while other pieces will involve detailed testing of the algorithm's response to different inputs. An example could be the automatic screening of new tenant applications in the rental housing market, where, e.g., any application scoring lower than X is automatically rejected. It is important to find out how the cutoff of X was found and justified. One would also want to test how slight changes in the cutoff differentially affect people belonging to different socio-economic groups.

The information collected about the algorithm and about the context will govern our assessments of the *key metrics*, which are *ethically salient features* of the algorithm in the relevant context. Table 1 shows a list of key metrics. A brief characterization of each metric follows the table.

When assessing how well a particular algorithm does on one of these metrics, there are a variety of methods that need to be utilized and a variety of ways the assessment can be presented. A narrative assessment (supplemented by possible numerical results of algorithmic testing) is the most informative, but numerical (e.g., a scale from 1 to 5) or categorical (bad, indifferent, good) assessments are also possible. All such assessments require standardized rubrics. Much more work should be done to lay out this part of audit in more detail.

Ideally, an auditor should be able to test an algorithm's performance and assess how well it is doing on each of the metrics independently of any of the other metrics and independently of stakeholder interests. By *independent* we mean that the metric is a measurable feature of an algorithm that does not depend strongly on other features, i.e., it can be objectively assessed in relative isolation of other parameters. For example, consider one of the metrics–*societal bias*. We could test for societal bias by looking to see whether individuals belonging to some particular societal group (race, gender, culture, economic status, etc.) are favored with systematically higher or lower scores by an algorithm; we could then assign a numerical or categorical score on the societal bias metric independent of whether this actually threatens any stakeholder interests.

Category	Metrics
Bias	Societal bias
	Statistical bias
Effectiveness	Accuracy
	Stability and repeatability
	Efficiency of data use
Transparency	Transparency of architecture
	Explainability
	Transparency of use
	Transparency of data use & collection
Direct Impacts	Potential for misuse and abuse
	Infringement of legal rights
Security & Access	Security & access in use of the algorithm
	Data security & access

Bias

Statistical bias. The formal definition for statistical bias is "the difference between an estimator's predicted value and the true value", and this can take many forms. Examples include simple offsets that are applied uniformly to all the outputs (e.g., everyone gets a lower credit score than they "should" according to some external standard), as well as systematic advantage for one or more privileged groups/classes stemming from a *statistical offset* in the training data (e.g., undersampling or oversampling data from particular groups/classes). There are cases where this undersampling/oversampling reflects an actual societal bias (see below). We should note that even biases that are uniformly applied across groups could be ethically significant depending on other external factors such as, for instance, if those groups have differential access to redress mechanisms.

Societal bias. A systematic advantage in the algorithm for one or more privileged groups/classes and/or a systematic disadvantage for one or more underprivileged groups. To be a societal bias, it must reflect a bias that exists within society and has been encoded either implicitly or explicitly, usually through the use of biased training data. As in statistical bias, undersampling or oversampling data from particular groups/classes can lead to differential outcomes, and would constitute societal bias in cases where this imbalance of data is societal in origin. An example is the fact that smartphone data oversamples those that can afford smartphones, and algorithms making use of this data may favor the preferences of those groups for which wealth is a proxy.

Effectiveness

Accuracy. A statistical measure of how "accurate" the algorithm is, and the measure itself will depend on the algorithm. The obvious case is a classification algorithm, in which case we measure how accurately the algorithm can correctly classify the input/subject. Analysis of accuracy might also include any discrepancies between effectiveness and promised effectiveness.

Stability & repeatability. A statistical measure of how robust the output is to minor/irrelevant variations in the input. This includes temporal variations (if we put the same input in at two different times, is the output the same?).

Efficiency of data use. This metric measures how efficiently the algorithm uses input data, both in terms of what data it takes in, and how it uses it within the algorithm. Typical questions one would ask to evaluate this are: Does the algorithm make use of irrelevant data? Is the algorithm making unnecessary computations with the data? Could a simpler (and more explainable) algorithm be used and still sufficiently fulfill the stated purpose?

Transparency

Transparency of architecture. A measure of how well the structure of the algorithm is known (or knowable) to stakeholders, including inputs and outputs. Is it a neural-network, equation, or a logical/semantic relationship? What are the weights of the network or the details of the equation?

Explainability & interpretability. For any given use of the algorithm, explainability and interpretability reflects how well one can know why and how a particular output was given.

Transparency of use. How transparent is the fact that the algorithm is being used? Examples would be online ad targeting or NSA surveillance (which until recently was opaque to most users) vs. credit scoring, which is widely known to be used.

Transparency of data use & collection. A measure of how well the collection and subsequent use (processing) of data for the algorithm is known to stakeholders. Important questions to answer include: Do users know what data about them is being collected, and how long it is being stored? Are stakeholders aware of what further processing or inferences will be made using the data, and for what purpose?

Direct impacts

Potential for abuse. A metric that measures the potential for the algorithm to be used to infringe on stakeholder rights or be used in other destructive/dangerous ways. By its nature, this metric requires that we look beyond the stated purpose/context and evaluate potential destructive uses of the algorithm. One of the key questions that context might help answer is what counts as misuse or abuse for a particular algorithm.

Infringement of (legal) rights. For cases where the very use of the algorithm in a particular context violates stakeholder rights (e.g., some autonomous weapons, explicit and illegal discrimination based on protected characteristics). As these issues (direct impacts) require consideration of stakeholders' interests and thinking beyond the stated context of the algorithm's use, they do not fit neatly into our framework. However, they are critically important for catching some of the most blatantly nefarious uses of algorithms.

Security & access. This category focuses on peripheral facts about who and what has the ability to use the algorithm and access the associated data.

Security & access of use. An assessment of how secure the use of the algorithm is. Who within an organization can use it? Can people outside the organization use it, and with what restrictions?

Data security & access. An assessment of how secure the data associated with the algorithm is. This includes data collected by the algorithm as well as the inferences produced as output.

Connecting the blocks: Relevancy matrix

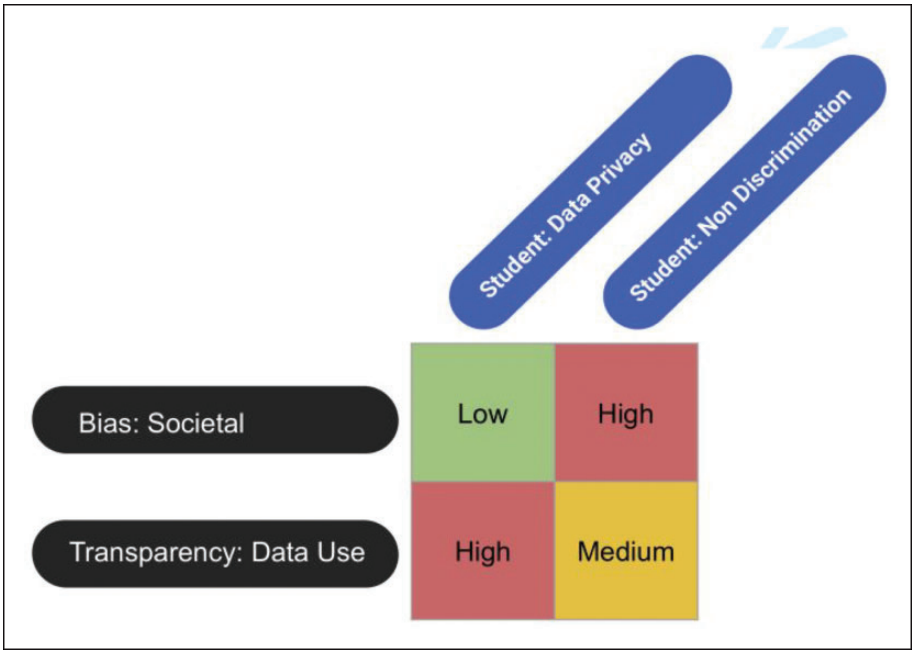
The general building blocks must somehow be used to fulfill the goal of the algorithm audit. The analysis of stakeholder interests (given the context) flags important interests that may come under threat, and the metric scoring highlights problematic features of the algorithm itself, but these two pieces need to be connected in a coherent way. One critical question needs to be answered in order to do this: *For each stakeholder interest, how relevant is each metric to the protection of that interest?* Stated another way: *For each stakeholder interest, how much could each metric threaten that interest if the algorithm performs poorly with respect to that metric?* As with the metrics, the way we present the answer to this question could take any form one would like (relevant/irrelevant, high/medium/low relevance, numerical score like 75% relevant, etc.). Answering this question for each interest and each metric forms a kind of matrix, which we call the *relevancy matrix*. Let us illustrate again with a fictitious AI essay-grading algorithm example, focusing on just two metrics and two interests for the student stakeholder: *privacy* and *non-discrimination* for the interests, and *societal bias* and *transparency of data use and collection* as the two metrics. Essay grading algorithms are widely used and provide automatic assessment of students' essays.

In our example (Figure 1), we have used the facts about the context of the algorithm to determine that a student's interest in data privacy is not significantly threatened by whether the algorithm shows bias based on socioeconomic factors (low relevance). However, the details of how much the student knows and consents to what happens to their data is highly relevant to privacy, independent of what those details actually are (which would be assessed in the metric score). For non-discrimination, societal bias is highly relevant, and for the purposes of this hypothetical example we score the transparency of data use and collection as medium relevance for non-discrimination. This stems from the possibility that student data is used to update the algorithm in some way, which could lead to compounding bias.

In a detailed audit, each of the elements of the matrix would be accompanied by a narrative justification of the stated relevance, and our choice of (high, medium, low) could be replaced by another scale deemed appropriate by a regulatory or standards organization. We should note that, as with our assessments of metrics, we would like the relevance to rely only on the context, and be as independent as possible from the outcomes of the stakeholder analysis and algorithm testing. This allows for efficiency of the process (different groups can work independently), reuse of analysis, comparison between algorithms that share similar contexts or working components (i.e., the same algorithm being used in a different context), and the potential to catch "double-counting" or circular reasoning in the overall analysis.

Audit results

Once one is armed with the relevancy matrix that connects the stakeholder interests to the metrics, the



	Student: Data Privacy	Student: Non Discrimination
Bias: Societal	Low	High
Transparency: Data Use	High	Medium

Figure 1. Example 2 × 2 relevancy matrix.

fulfillment of the audit goal is straightforward, though it can proceed in a variety of ways. The most obvious is to identify low scoring metrics that have high relevance to stakeholder interests. This will flag areas of potential negative impact, and the narrative assessment of the metric testing can be used to highlight strategies to mitigate this risk. Some industries or organizations where algorithms have a very narrow focus (e.g., credit scoring, tenant screening by property managers, etc.) might prefer to develop a more numerical approach, where algorithms can be rated and compared to each other based on how well they perform in certain metrics. The result of each audit will be a two-value (qualitative or quantitative) score, with the first value describing or scoring the metric and the second describing or scoring the relevance of that metric to some interest (i.e., interest: metric score, relevancy score). The purpose for which an audit is done (regulatory, risk management, general ethical assessment) should inform how the audit result/score is presented. For those interested in the audit for regulatory purposes, it would be sufficient to first identify stakeholder interests that the regulatory agency legitimately regulates and then for each of those interests examine whether there are any cases where the algorithm performs low on some metric that is highly relevant (key interest: low, high). For those interested in risk management any (low, high) score for any interest will require a detailed explanation that could provide a way to mitigate the reputational, ethical, or financial risk that a poorly performing algorithm might present. For those interested in what we called a general ethical assessment, performing poorly on almost any metric that is important for the interests the user of audit finds relevant will be reason to reject or protest the use of such an algorithm. It is important to note that recently much criticism has been directed at early attempts to provide ethical analysis of algorithms. Scholars have argued that using the classical analytic approach that over-stresses technical aspects of algorithms and ignores the larger socio-technical power dynamics has resulted in ethical approaches to algorithms that ignore or marginalize some of the primary threats that (especially decision-making and classification) algorithms pose to minorities (Benjamin, 2019; Cave and Dihal, 2020; D'Ignazio and Klein, 2020; Mohamed et al., 2020). We share this view, and think significantly more work needs to be done to gain additional insights from critical approaches to ethical analysis of algorithms. Our highly context-dependent approach to audits is meant to be sensitive to these worries while staying within the constraints of what a genuine audit can do, which is to provide consistent and repeatable assessment of (in this case) algorithms.

Acknowledgments

We wish to thank Mihailis Diamantis, Marco Meyer, Mitt Regan, Ron Sandler, three anonymous referees and the editors for this journal, Prof. Davidovic's Ethics of Algorithms graduate seminar students, and the audience at the 17th Annual Information Ethics at the Northeastern University for feedback on this article.

Declaration of conflicting interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jovana Davidovic https://orcid.org/0000-0002-8998-5496
Ali Hasan https://orcid.org/0000-0003-2963-2573

References

Ada Lovelace Institute (2020) Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf

Barocas S, Hood S and Ziewitz M (2013) Governing Algorithms: A Provocation Piece. Available at: http://dx.doi.org/10.2139/ssrn.2245322 (accessed 28 November 2020).

Benjamin R (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. New York, NY: John Wiley & Sons.

Brundage M, et al. (2020) Available at: https://arxiv.org/abs/2004.07213

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *2019 ACM conference on fairness, accountability, and transparency (FAT*)*, 23–24 February 2018.

Cave S and Dihal K (2020) The whiteness of AI. *Philosophy. & Technology* 33: 685–703.

Deville J (2019) Digital subprime: Tracking the credit trackers. In*: The Sociology of Debt*. Chapter 6. Bristol: Policy Press.

D'Ignazio C and Klein LF (2020) *Data Feminism*. Cambridge, MA: MIT Press.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St Martin's Press.

Floridi L, et al. (2018) AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach (Dordr)* 28: 589–707.

Future of Privacy Forum (2017) Unfairness by algorithm: Distilling the harms of automated decision-making. Available at: https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/ (accessed 28 November 2020).

Gebru T, et al. (2018) Datasheets for datasets. Available at: https://arxiv.org/abs/1803.09010

Mitchell M, et al. (2019). Model cards for model reporting. In: *Proceedings of the conference on fairness, accountability, and transparency (FAT* '19)*, Association for Computing Machinery, New York, NY, USA, pp. 220–229. Available at: https://doi.org/10.1145/3287560.3287596

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1: 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3.

Mohamed S, Png M, Isaac W, et al. (2020) Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33: 659–684.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Obermeyer Z, Powers B, Vogeli C, et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.)* 366(6464): 447–453.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Broadway Books.

Prabhu VU and Birhane A (2020) Large image datasets: A pyrrhic win for computer vision? *arXiv preprint* arXiv:2006.16923.

Raji ID, et al. (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (FAT* '20), Association for Computing Machinery, New York, NY, USA, pp. 33–44. Available at: https://doi.org/10.1145/3351095.3372873

Raji ID and Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society (AIES '19),* Association for Computing Machinery, New York, NY, USA, pp. 429–435,.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* 22.

Schiff D, et al. (2020) Available at: https://arxiv.org/abs/2006.04707

Selbst AD, Boyd D, Friedler S, et al. (2018) Fairness and abstraction in sociotechnical systems. In: *2019 ACM conference on fairness, accountability, and transparency (FAT*)*, 23–24 February 2018, pp. 59–68 Available at: https://ssrn.com/abstract = 3265913 (accessed 28 November 2020).

Whittaker M, Crawford K, Dobbe R, et al. (2018) *AI now report 2018*. AI Now Institute, USA. Available at: https://ainowinstitute.org/AI_Now_2018_Report.pdf