

The politics and reciprocal (re)configuration of accountability and fairness in data-driven education

Otto Sahlgren

To cite this article: Otto Sahlgren (2021): The politics and reciprocal (re)configuration of accountability and fairness in data-driven education, Learning, Media and Technology, DOI: [10.1080/17439884.2021.1986065](https://doi.org/10.1080/17439884.2021.1986065)

To link to this article: <https://doi.org/10.1080/17439884.2021.1986065>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 01 Oct 2021.



[Submit your article to this journal](#)



Article views: 732




[View related articles](#)



[View Crossmark data](#)

The politics and reciprocal (re)configuration of accountability and fairness in data-driven education

Otto Sahlgren 

Faculty of Social Sciences, Tampere University, Tampere, Finland

ABSTRACT

As awareness of bias in educational machine learning applications increases, accountability for technologies and their impact on educational equality is becoming an increasingly important constituent of ethical conduct and accountability in education. This article critically examines the relationship between so-called algorithmic fairness and algorithmic accountability in education. I argue that operative political meanings of accountability and fairness are constructed, operationalized, and reciprocally configured in the performance of algorithmic accountability in education. Tools for measuring forms of unwanted bias in machine learning systems, and technical fixes for mitigating them, are value-laden and may conceal the politics behind quantifying educational inequality. Crucially, some approaches may also disincentivize systemic reforms for substantive equality in education in the name of accountability.

ARTICLE HISTORY

Received 19 May 2021

Accepted 22 September 2021

KEYWORDS

Algorithmic fairness; accountability; education policy; ethics; equality

1. Introduction

Data-driven education and educational artificial intelligence technologies (AIED) promise ‘flexible, inclusive, personalized, engaging, and effective’ teaching and learning (Luckin et al. 2016, 18), and increased equity and accountability (see Fontaine [2016]). However, AIED technologies themselves can involve significant risks of algorithmic discrimination and unfairness in pedagogical and administrative practices with high-stakes and which are conducted in educational domains already characterized by patterned inequalities and structural discrimination (Baker & Hawn [2021]; Kizilcec & Lee [forthcoming]). Machine learning (ML) systems, which dig through troves of student and learning data, learn biases due to technical design errors but also by picking up on existing patterns of inequality and discrimination (Baker and Hawn 2021).

Under increasingly closer scrutiny, schools and policymakers search for ways to prove that their algorithms are effective, equitable, and trustworthy (see Zeide [2020]). What is called *algorithmic accountability* (Binns 2018; Reisman et al. 2018) thereby emerges as a nascent constituent of educational institutions’ ethical conduct and accountability. Due to increased awareness of problems with algorithmic bias, *algorithmic fairness*, in turn, comes to constitute a key aspect of algorithmic accountability (Kizilcec and Lee forthcoming), creating a market for instruments applicable for legitimating AIED systems’ equitable impact. Research in fair ML, a subfield of technology ethics, has developed various measures for detecting wrongful, discriminatory biases in ML systems, and technical fixes for mitigating them. Although sharing in underlying motivations, these measures reflect different ethical views about fairness and equality (Narayanan 2018). No consensus on the proper measure of algorithmic fairness has emerged. Rather, what algorithmic fairness concretely

CONTACT Otto Sahlgren  otto.sahlgren@tuni.fi

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

entails, and for whom, is debated in the ‘terrain of contestation’ that is technology ethics (Green 2021).

This article critically examines resources developed in fair ML research as emerging instruments for performing algorithmic accountability in education. Drawing on critical work on the politics of educational accountability and accountability instruments, and technology ethics, I argue that algorithmic fairness as performance constructs operative political meanings of fairness and equality in data-driven education. This, in turn, (re)configures accountability relationships by demarcating which inequalities educational institutions are responsible for mitigating and how. Crucially, I argue that algorithmic fairness may incentivize narrow interventions while legitimizing and depoliticizing substantive inequality. The substantive question regarding the demands of fairness is beyond the present scope. Rather, I will consider a set of intertwining epistemological and political questions: how and why do algorithmic fairness analyses construct meanings of fairness and inequality, and how does this configure accountability relationships between educational institutions and the public?

The article is structured as follows. Section 1 discusses the emerging role of algorithmic accountability in education against the broader frame of educational accountability and instrumentation therein. Legitimation and the political performance of accountability with instruments and data is employed as a critical theoretical lens throughout the section. Section 2 examines fair ML resources as instruments for performing algorithmic accountability, focusing on fairness metrics, in particular. The section demonstrates how fairness analyses construct political operative meanings of (in)equality and (un)fairness in education, and how they involve (and may conceal) legally and morally significant decision-making and dynamics of inclusion and exclusion. Section 3 argues that certain approaches to algorithmic fairness risk legitimizing existing structural inequalities and may disincentivize reforms that would promote substantive equality, all in the name of accountability. The final section summarizes the study’s findings.

2. (Algorithmic) accountability and its politics

Accountability can be characterized as ‘a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences’ (Bovens 2007, 450; italics removed). Specifically, discussions on educational accountability commonly revolve around performance reporting, comparative assessments and (self-)evaluations; schools’ effective and equitable delivery of the services they produce; and questions of political legitimacy and stakeholder fairness (see Levin [1974]; Fontaine [2016]). Increasingly, so-called high-stakes accountability, in particular, is characterized by a technical-managerial focus on compliance and conformity to principles of good governance through the use of standardized and data-intensive policy instruments, as well as neoliberal market logics (see Diamond and Spillane [2004]; Fontaine [2016]; Ozga [2013]; Ozga [2019]; Verger, Fontdevila, and Parcerisa [2019]). Indeed, various instruments (e.g., school performance tables, test-based accountability systems, and (inter)national large-scale assessments such as PISA studies) are used to produce and disseminate information about public sector performance across networks of actors, such as educational institutions, regulatory bodies, and localities. In decentralized education and governance, accountability systems and policies are taken to render schools into ‘service providers who must navigate a competitive choice marketplace as they vie for students/customers’, which in turn requires ‘distinguishing their market niche and by using data to demonstrate their effectiveness’ (Fontaine 2016, 8). Meanwhile, the demand for data-for-accountability has created, and continues to consolidate, a market for data-based instruments competing for the status of ‘best evidence’ (Ozga 2019, 3–4; Williamson and Piattoeva 2019). Winners in this market are instruments that generate ‘simplified, comparative, normative, [and] transportable’ knowledge (Ozga 2019, 8) which can be produced locally and compared globally.

With digital reform and increasing procurement and implementation of AIED systems in educational domains, a novel socio-technical layer emerges into accountability relationships: accountability for automated tools, predictive algorithms, and their impact (see Zeide [2020]; Binns [2018]). I will next discuss algorithmic accountability in the context of broader discussions on educational accountability, accountability instruments, and their politics.

2.1. Accountability with and for algorithms in education

Data-driven decision-making promises more effective, equitable, and accountable pedagogy and administration at both smaller and larger scales (Luckin et al. 2016; Fontaine 2016, 3). The dominant vision is that the use of massive amounts of data combined with the processing capabilities of automated systems weeds out arbitrariness, human flaws and bias, meanwhile offering actionable insight. This tendency to privilege ‘quantification and statistical analysis as ways of knowing’ is, notably, characteristic to accountability discourse, more generally (Fontaine 2016, 2). As statistical knowledge invokes a sense of objectivity (Desrosières 2001; Williamson and Piattoeva 2019), it contributes to the (re)production of political legitimacy of educational institutions by evincing the effectiveness and equitability of their services through numbers.

Visions of equitable and accountable data-driven education are being questioned, however, due to emerging concern over ‘algorithmic bias’. This is evinced by the now well-known case where Ofqual’s A-levels exam grade standardization algorithm downgraded almost 40 per cent of students’ grades, primarily disadvantaging Black, Asian, and Ethnic Minority students and students of low socio-economic status (Wachter, Mittelstadt, and Russell 2021). The public demanded accountability after what can only be described as a fraught and messy policymaking process involving competing conceptions of fairness, unavoidable harms, and an insufficient appeal system (Adams, Weale, and Barr 2020; Kippin and Cairney 2021). Algorithmic bias raises concern in various application areas of AIED technologies, including educational assessment, students’ dropout risk prediction, and algorithmic ability-grouping (see Baker & Hawn [2021]; Kizilcec & Lee [forthcoming]; Carmel & Ben-Shahar [2017]).

‘Algorithmic accountability’ is envisioned to address ethical challenges with AIED. Education scholarship has called for ‘rigorous and documented procedures for edtech procurement, implementation, and evaluation’, ‘sufficient access to underlying data and information about outcomes’, and ‘mechanisms for transparency about individual decisions and access to outcomes for relevant populations and protected classes to enable ongoing review’ (Zeide 2020, 802). Best practices, mechanisms, and instruments for algorithmic accountability are widely debated in the field of technology ethics (Green 2021), however, complicating decision-makers’ navigation in the market for applicable instruments. These debates provide context for the critical examination of algorithmic fairness and its relation to (algorithmic) accountability in education. First, however, I will describe the conceptual frame of the study.

2.2. Conceptual frame

This article takes accountability and its performance through practices and policies as political *qua* (re)produced in multiple national and global contexts among competing interests, relationships of power, discourses, and social contingencies (Ozga 2019; Lascoumes and Le Galès 2007). Accountability relationships are built and consolidated in networks of governance; continuously constructed, distributed, and stabilized through various practices of knowledge production and dissemination, evaluation, and performative practices by public agents (Ozga [2013]; Ozga [2019]). Accountability policies are typically the result of multiple stages of normative and contestable decision-making, including social processes of conceptualization, operationalization, implementation, and evaluation (Ozga 2019, 2–3).

Instruments used to perform accountability are likewise malleable and contestable, and their implementation contingent on various factors. Research on instrumentation has shown that there are diverse reasons, rationales, and discourses (political, economic, and institutional) that condition the adoption and selection of policy instruments (Lascoumes and Le Galès 2007; Capano and Lippi 2017). Though policy instruments (e.g., national large-scale assessments and test-based accountability systems, such as PISA) have been globally adopted, (motivations for) their actual uses are ‘contingent to the specificities of the political and institutional settings where they are embedded’ (Verger, Fontdevila, and Parcerisa 2019, 249).

For purposes of clarity, I distinguish between so-called *first-order* and *second-order accountability instruments*. Algorithmic systems and standardized tests, for example, function as both education instruments and first-order accountability instruments. They can be primarily used for pedagogical purposes (e.g., to generate information for assessment, to predict student outcomes), but they also help agents perform accountability by consistently generating information for secondary purposes (e.g., for teacher evaluation or school performance comparison). These instruments are themselves to be accounted for, however: information about their impact is produced to assess them and to legitimate their use. Fairness measures for tests and algorithms, for example, have similar “second-order” functions, accordingly. Performance of algorithmic accountability is defined here as the second-order use of evidence regarding and/or produced by algorithmic systems used for first-order purposes, respectively. What distinguishes AIED systems from traditional tests, however, is the fact that confidentiality and proprietary nature of software and models may preclude access to evidence on their fairness (The Institute for Ethical AI in Education 2021; Digital Promise 2014). In contrast to tests *qua* predictive instruments, some AIED systems also update their models continuously based on new data.

Now, a key notion in this study is that knowledge about accountability relationships, and operative social and political meanings therein, can be gained by examining how they are constructed and performed with instruments (Lascoumes and Le Galès 2007). As ‘every instrument constitutes a condensed form of knowledge about social control and ways of exercising it’ (Ibid., 1), dissection of accountability instruments can reveal politically significant assumptions behind the meanings and relations (re)constructed through instrumentation itself. The data produced with them have various functions: they make the invisible visible to the public through mechanisms of transparency; they open up ways for networks of actors to understand practices, activities, and themselves; and they position actors within those networks and relations of power (Piattoeva 2015). Depoliticized conceptions of instrumentation persist, however, as ‘instruments of data collection and analysis [...] invoke scientific and technical rationality, which combine to support their legitimacy and neutralize their political significance’ (Ozga 2019, 5). The dominant statistical way of seeing, a sort of metric realism (Desrosières 2001), is taken to capture phenomena as if independent from observation and the processes that (re)produce them. Consequently, the politics of instrumentation are rendered invisible, enabling ‘a technocratic elite to accrue power and influence in education, to profit from its provision and to pursue political agendas that appear to be neutral, objective and necessary’ (Ozga 2019, 9).

Another key notion of this study is that operative forms of accountability and instruments used to perform them mutually configure each other. For example, regard for accountability and good governance configures concrete practices of data collection, interpretation, and use, as is broadly exemplified by the datafication of education (Jarke and Breiter 2019; Sellar 2015). Local effects can be found in how schools mobilize data for high-stakes accountability purposes, using data to identify “broken parts” in educational practices and to bring learners on the “curb of proficiency” above thresholds specified by accountability policies (Diamond and Spillane 2004; Datnow and Park 2018).

However, accountability instruments also actively (re)configure operative meanings of accountability itself, alongside adjacent activities. Metrics, indicators, and data create and distribute meanings, and legitimate action, both constructing policy problems and framing solutions to them (Ozga

2019, 4). The ‘instruments at work are not neutral devices: they produce specific effects, independently of the objective pursued’ (Lascoumes and Le Galès 2007, 1). For example, a study on fairness measures in German school monitoring and management showed that

once particular indicators [...] become ‘fixed’ and thus objectified as representations of inequality, a growing number of continuous monitoring and management practices may become affected by these indicators, such as practices of school performance comparison or accountability measures. [...] [A]s schools become increasingly surveilled through the use of (meta)data [...] social indices can become a powerful form of such (meta)-data, which ultimately inform multiple sorting, classification and ordering practices, and thus stabilize the (dated) inequality between schools. (Hartong and Breiter 2020, 58.)

Pressures relating to the demands of accountability can also lead to undesirable "ripple effects", such as efforts to "game the system" through data manipulation (Fontaine 2016, 5–6).

In other words, accountability instruments not only produce knowledge for accountants’ needs but they also actively construct the measured phenomenon itself, thereby orienting future practice. They configure, but are also configured by, existing relationships between actors, relations of power, and social practices. Drawing on these notions, I maintain that algorithmic accountability instruments constitute technologies of governance and control, and instruments for the (re)production of political legitimacy. They open and expose ‘black box’ algorithms, their effects, and limitations to stakeholders (e.g., users, citizens, and regulatory bodies) to generate trust in governance and its technologies. However, the data produced by and with these instruments effectively gain and exercise power in processes of governing algorithms, positioning actors, and configuring networks of responsibility. I call this dynamic interplay between, and co-construction, of operative meanings of fairness, on the one hand, and socio-technical practices that perform accountability, on the other, the *reciprocal (re)configuration of accountability and fairness*.

3. Algorithmic fairness for accountability?

Trained on historical data, ML algorithms may infer (multiple) proxies for legally protected or otherwise sensitive attributes (e.g., ‘race’, gender, or socio-economic status), consequently introducing disparities into algorithmic predictions (Baker & Hawn [2021]; Mehrabi et al. [2021]). Due to unrepresentative sampling and other technical design flaws, bias against specific groups in outcomes may occur without explicit use of protected or sensitive attributes as model features. Notably, when algorithms are optimized for predictive accuracy on structurally biased historical data, they can effectively reproduce inequalities captured by those data even in the absence of technical bias (Wachter, Mittelstadt, and Russell 2021).

Educators and policymakers increasingly search for ways to put biased algorithms in check. Public agencies are called to conduct self-assessments of ‘existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities’ (Reisman et al. 2018, 4). Evidence of algorithmic fairness (e.g., lack of data or model bias) is understood as essential when schools are looking to procure AIED technologies from private vendors (Institute for Ethical AI in Education 2021; Zeide 2020). What exactly constitutes wrongful bias is under debate, however, and research in fair ML provides competing answers in this respect. The field has developed various frameworks and technical tools for ensuring fair treatment and outcomes in the use of ML systems, including over 20 formal definitions for fairness translated into metrics, respectively (Narayanan 2018). Diversity in competing interpretations regarding "ethics" and its practical demands is, notably, characteristic of the broader field of technology ethics as well (Green 2021).

Fair ML metrics bear similarities to fairness measures for tests in education (see Camilli [2013]; Hutchinson & Mitchell [2019])¹, and they have been increasingly discussed in educational contexts (Kizilcec & Lee [forthcoming]; see also Baker & Hawn [2021]). They have a two-fold function: Firstly, designed to capture distinct philosophical and legal concepts of discrimination and equality, they function as idealized formal measures for fairness. Secondly, by indicating deviations from the

ideal (i.e., unwanted bias), they inform decision-makers about the need for interventions. Existing methods can mitigate identified biases by transforming the composition of the data used to train the algorithm, by adjusting the training process or the algorithm, or by balancing the output distribution (see Mehrabi et al. [2021]). Open-source toolkits (e.g., Aequitas² and AI Fairness 360³) enable developers of ML systems to evaluate their pre-trained models against a variety of fairness definitions, aiding in efforts to mitigate bias. These and other similar toolkits are envisioned to be used ‘by developers of algorithmic systems in education for internal audits and by education policy-makers for external and periodic audits’ (Kizilcec and Lee forthcoming, 19).

Most fairness definitions balance some statistical metric in the predictive model (e.g., predicted positive outcomes or error rates) across comparison classes. Given the abundance of definitions, I mention only a few notable ones to structure the discussion (Table 1).

To illustrate, consider university admissions. *Statistical Parity* is satisfied when members of the compared protected groups are equally likely to be predicted to succeed in their studies (i.e., admitted). When the probability scores (e.g., predicted chances of succeeding) generated by an algorithm accurately estimate underlying population values (e.g., actual likelihood of succeeding) the algorithm is *Calibrated*. *Equalized Odds* calls for equal error rates: applicants from different protected groups should be equally likely to receive the wrong decision given their actual chances of succeeding. In addition to these “group fairness” metrics, *Fairness Through Awareness* requires that across all pairs of applicants, the generated probability scores differ only to the extent that those individuals differ from each other in their non-protected attributes (e.g., academic attainments). The stakes of ongoing debates regarding algorithmic fairness have been raised due to discoveries of trade-offs: fairness typically comes with a cost to predictive accuracy (Dwork et al. 2012), and algorithms cannot satisfy certain fairness definitions simultaneously, except in highly unlikely circumstances (Kleinberg, Mullainathan, and Raghavan 2017). (See also Camilli [2013] on test fairness.)

Now, while the mantra “technology is political” is repeated *ad nauseam*, we would do well to note that there is also a politics to fair ML resources *qua* algorithmic accountability instruments. Abstraction, interpretation, sense-making, and value-laden decisions are involved in fairness analyses, which have legal and ethical implications when they guide interventions and (re)configure practice. To place ‘algorithmic fairness’ under critical scrutiny as an accountability instrument and a rhetorical device in educational contexts, I will examine fairness analyses and the methodological frame of fair ML to highlight stages of normative decision-making, and dynamics of inclusion and exclusion.

3.1. Constructing fairness

Building an algorithm for a given purpose involves translation of abstract goals and objectives into computationally tractable problems. Algorithmic fairness analyses involve similar work. Methodological choices therein effectively construct an operationalized conception of (in)equality at stake

Table 1. *Fairness definitions.* A binary prediction/decision is denoted with $D = \{0, 1\}$ and the actual outcomes are denoted with $Y = \{0, 1\}$. A probability score $s \in S$ determines the value of D for each individual based on a certain decision threshold in the predictive model. Model features are denoted with V . A set of non-protected attributes is denoted with $X \in V$. Membership of a protected group is denoted with $G = \{a, b\}$.

Definition	Description	Formal definition
<i>Fairness Through Unawareness</i>	Model features do not include protected attributes.	$G \notin V$
<i>Statistical Parity</i> (e.g., Dwork et al. 2012)	Members of protected groups have equal probability of belonging to the class of predicted positive outcomes.	$P(D = 1 a) = P(D = 1 b)$
<i>Calibration</i> (Chouldechova 2017)	Probability scores estimate underlying population values accurately.	$P(Y = 1 s, a) = P(Y = 1 s, b)$
<i>Equalized Odds</i> (Hardt, Price, and Srebro 2016)	Protected groups have equal error rates (i.e., false negative rates and false positive rates).	$P(D = 1 Y, a) = P(D = 1 Y, b)$
<i>Fairness Through Awareness</i> (Dwork et al. 2012)	Similar data profiles are be treated similarly.	[Requires defining a similarity metric, e.g., Euclidian distance]

which, in turn, focuses possible interventions and structures future practice. I will substantiate this issue by considering a variety of such decisions. Of course, in practice, the relevant decisions can be constrained by legislation (e.g., non-discrimination law), decision-makers' access to data, among other factors. The insights offered here are valuable regardless.

3.1.1. *Selecting and defining comparison classes*

In practice, algorithms cannot be audited for fairness across all possible (sub)groups, meaning certain groups must be prioritized in selecting comparison classes for fairness analyses. Protected groups specified in non-discrimination legislation (e.g., 'race', gender) are common choices, although other factors (e.g., urbanicity, socio-economic status) may be accounted for (Baker and Hawn 2021). The results of fairness analyses depend on what categories are selected as well as on how they are constructed – e.g., how 'racial categories' are abstracted 'into a mathematically comparable form' (Hanna et al. 2020, 508). Variance in these respects can affect the comparability of information produced with fairness analyses across different educational institutions and contexts.

3.1.2. *Selecting a fairness metric*

The objectivity of policy instruments is produced via a threefold translation: first, translation of 'scientific expertise into standardized and enumerable definitions'; second, translation of these standards into 'concrete technologies of measurement'; and third, translation of the 'data produced through measurement technologies into objective policy-relevant knowledge' (Williamson and Piattoeva 2019, 64). Likewise, fairness metrics comprise reductive, allegedly domain-neutral formulae which translate legal and ethical expertise into the language of algorithms. They are represented as computational implementations of stabilized anti-discrimination norms, which in turn figures into the social production of their objectivity and legitimacy. The relevant norms might also include *de facto* domain standards, such as test fairness measures in education. Standardized tests, for example, are evaluated for fairness by examining whether they 'produce the same score for two test-takers who are of the same proficiency level' (Hamilton, Stecher, and Klein 2002, xvi) and whether their predictive accuracy differs across groups (Ibid., 69–70). The first notion roughly corresponds to *Fairness Through Awareness* and the latter most closely to *Calibration*. On a broader understanding of model accuracy, including also false predictions, the latter requirement reflects *Equalized Odds*.⁴

Fairness definitions are not neutral among competing interests and ethical conceptions of fairness (Narayanan 2018). Metrics that neglect actual outcomes, such as *Statistical Parity*, can be met even when systems' predictions systematically err (Dwork et al. 2012), effectively uncoupling distributive fairness from perceived 'merit'. Fairness as *Calibration* (Chouldechova 2017) reflects one notion of formal equality; the same rules should apply to everyone regardless of their protected attributes. As a calibrated model's predictions mean the same thing across compared groups and probability scores, it makes the system more trustworthy from the perspective of the user. *Equalized Odds* (Hardt, Price, and Srebro 2016), in turn, prioritizes balance across groups in terms of risks for misprediction: fairness is understood as equal risk for misallocation or misrecognition of 'merit' or 'need' (Kizilcec and Lee forthcoming).

3.1.3. *Legitimizing disparities*

A predictive algorithm that treats everyone equally in the strictest sense fails to discriminate between the targets, i.e., attributes of interest. Accordingly, fairness measurements are typically conditioned on some set of model attributes. For example, *Statistical Parity* can be conditioned on a set of "legitimate" variables allowed to introduce deviations from a strict balance in groups' representation in the predicted outcome class (*Conditional Statistical Parity* in Corbett-Davies et al. [2017]). Some metrics, such as *Calibration* and *Equalized Odds*, condition the fairness measure on actual outcomes in the data. Drawing lines between "legitimate" and "illegitimate" proxies for protected

attributes in predictive models, decision-makers theorize their responsibilities regarding educational (in)equality. Specifying "legitimate" sources of disparity in outcomes, they effectively draw a line between what learners themselves are to be held accountable for, and what (possibly latent) sources of inequality decision-makers ought to address. A stance on what counts as a matter of 'race', gender, or disability, etc., is taken. The relevant decisions echo both a fundamental question of egalitarianism (what kinds of attributes may rightfully introduce deviations from strict equality of outcome, if any?) and related debates concerning the scope of schools' responsibilities in mitigating the influence of learners' childhood environments on their future prospects. Even if schools have a minimal duty to mitigate the 'unequalizing impact' of the former on the latter (Coleman 1975, 29), distinguishing legitimate "effort", "merit" or "ability" from latent socio-economic variables (e.g., environmental stressors and access to learning resources) that covary with protected attributes remains a complex moral question.

3.1.4. Trade-offs

Given well-known "achievement gaps" between racial and gender groups, trade-offs are likely to occur when attainment data is used in algorithmic decision-making in educational domains. Optimizing algorithms for predictive accuracy on structurally biased data prioritizes model accuracy over group fairness whereas increasing the latter can also reduce the former (Dwork et al. 2012). As mentioned, another trade-off is found between *Calibration* and *Equalized Odds*: when target attribute prevalence differs across groups, both metrics cannot be satisfied simultaneously, except in highly constrained cases (Kleinberg, Mullainathan, and Raghavan 2017). An apparent dilemma follows: either "level down" predictive accuracy or sacrifice fairness. To the extent that this is an actual dilemma (see Green & Hu [2018]), decision-makers must prioritize certain values over others, bringing questions of power and public accountability ever more strongly into the picture.

3.2. Abstraction in algorithmic fairness

When concepts from ethics and law are translated into algorithmic logics, contestable concepts are 'stabilized as standardized and enumerable categories that can be transported to new sites of practice and coded into new assessment services, metrics, and technology products', to quote Williamson and Piattoeva (2019, 73). Things are always lost in translation, and abstractions draw boundaries between alternative ways of knowing. When it comes to matters of equality, this can have legally and morally significant consequences. Critics have noted that due to abstraction of empirical details, dynamics, complexity, and deep socio-ethical issues, fair ML frameworks may lead to 'ineffective, inaccurate, and sometimes dangerously misguided' interventions (Selbst et al. 2019, 59; Green & Hu 2018; Green 2020). Alas, abstractions characteristic to formalist fair ML methodologies can be problematic irrespective of (or alongside) particular problems pertaining to specific metrics. I will highlight these issues by using algorithmic ability-grouping as an example (see Carmel & Ben-Shahar [2017]).

In ability-grouping, attainment data are used to predict students' success and to place them into "low" or "advanced" groups. Predictions have different benefits and costs here, the relative significance of which bear on considerations of fairness. For example, accurate predictions place students into groups according to perceived "merit". A false positive places a student into an ability-group where they are likely to fail, while a false negative unduly denies one access to a suitable group. The costs of errors fall on the students, but also on the teachers who bear the burden of accommodating for incorrect placement decisions in their teaching. When errors are unavoidable, prioritization decisions need to be made with due regard for these costs. However, when fairness is assessed only in terms of statistical metrics, realized outcomes and their relative significance are neglected by default. Realization of outcomes is also mediated by social factors typically uncaptured by fair ML frameworks (Selbst et al. 2019), such as whether algorithms' predictions are enacted by human users in the first place. In fair ML, model outputs are assumed to smoothly translate into actual outcomes,

but this is plausibly contingent on things such as system interpretability, and users' competence and bias (Selbst et al. 2019; Kizilcec & Lee [forthcoming]). Furthermore, when considering fairness in ability-grouping, what happens *within* the composed groups can be relevant. Say, if lower groups receive less resources or lower quality teaching than higher groups, assignment into a lower group could be perceived as unfair regardless of whether placement decisions are accurate (statistically speaking). Indeed, research suggests that relevant effects on learning and equity are constituted not by the practice of grouping as such, but by contextual factors and processes that take place within schools, and the ability-groups themselves (see Boaler, Wiliam, and Brown [2000]; Eder [1981]; Felouzis & Charmillot [2013]).

Lastly, fair ML frameworks are primarily concerned with parity and comparative justice. Consequently, they have largely failed to account for *non-comparative justice*, i.e., what students are owed independently of one another, and what means of transparency and redress are needed, for example (Selbst et al. 2019). These considerations are, however, increasingly significant as opaque "black box" algorithms leave few possibilities for students and their families to understand or contest automated decisions (Zeide 2020).

3.3. Portability aspirations

Acceptance of, and reliance on, accountability instruments and other policy tools are created in networks of power relations and agents who require 'simplified, comparative, normative, [and] transportable' knowledge (Ozga 2019, 8). Given the reliance on local self-evaluations in decentralized educational governance, on the one hand, and the supply of AIED systems coming primarily from private vendors, on the other, transportability is demanded also from algorithmic accountability instruments. Fair ML frameworks and toolkits are attractive in this respect as, reflecting the aspirations of portable design characteristic to ML products and software more generally, they are often proposed as universally applicable across applications (see Green & Hu [2018]; Selbst et al. [2019]). In contrast, research on test fairness in education has acknowledged the politics of fairness metrics, recognizing the possibility that no universal agreement on fairness may come to fruition (see Camilli [2013]; Hutchinson & Mitchell [2019]).

The portability aspirations of fair ML prove problematic, firstly, because they render considerations of fairness insensitive to local concerns, variations in contexts and populations, and other sociotechnical factors discussed above. A given algorithm benchmarked as fair may not lead to similar realizations of social outcomes in different educational settings due to such variance; 'the local fairness concerns may be different enough that systems do not transfer well between them' (Selbst et al. 2019, 61). Secondly, due to the opaque, asymmetric relationships between private vendors of AIED systems and agents in the educational domain, the latter can be reliant on assessments conducted by the former, whose methodologies for bias identification and mitigation, and prioritization decisions regarding the eventual model design, may not be disclosed. So-called 'fairness washing' creates further concerns in this regard. To increase ML products' attractiveness in the market, private vendors may conduct fairness analyses against metrics of their choosing, allowing them 'to claim a certificate of fairness' while simultaneously shielding them from claims of immoral conduct as they continue 'weaponizing half-baked [automated] tools' (Fazelpour and Lipton 2020, 61).

These issues pose a challenge for both the transportability of fair ML frameworks as algorithmic accountability instruments, and the credibility of private vendors' fairness-certified software and their off-the-shelf use across educational settings. The question regarding the degree of influence private vendors have in these respects is deserving of its own study, nonetheless.

4. "Fixing" inequality with algorithmic fairness

The previous section demonstrated that algorithmic fairness analyses involve epistemological and ethical assumptions, and abstractions, which can have practical consequences when AIED systems

are used at scale. This section will further illuminate the issue. Recall that high-stakes accountability-driven practices of data collection, interpretation, and use in schools focus on identifying problems, comparative performance evaluations, and administrative compliance. While data-driven technologies are ‘imagined to create new forms of accountability’ in these respects, the discourse shows ‘little consideration for the longstanding history of current trends’ that questions the link between data utilization and promotion of equality (Fontaine 2016, 3). I argue that algorithmic accountability can suffer from a similar second-order neglect: without an explicit aim to promote substantive equality, “algorithmic fairness” can risk discouraging non-technical reforms in educational settings and systems. Paradoxically, technical fixes may end up *fixing* inequalities only in the sense of solidifying them through political processes of legitimation.

4.1. Technical fixes for broken parts

To boost achievement metrics, schools are keen on bringing learners on the “curb of proficiency” above a certain threshold; to fix broken parts, as opposed to conducting extensive reforms which would benefit all students (Diamond and Spillane 2004; Datnow and Park 2018). With its focus on technical as opposed to *socio*-technical systems, fair ML can promote interventions exemplary of this “broken part fallacy”. For one, certain technical interventions involve changing the prediction labels (e.g., from “fail” to “pass”) of disadvantaged groups’ members close to the decision-threshold, quite literally on the “curb of proficiency”, to increase model fairness (Kamiran, Karim, and Zhang 2012). More generally, however, fair ML methods focus on fixing technology as opposed to social structures and practices, thereby incentivizing narrow technical interventions (e.g., data collection, model adjustment) ‘at the risk of missing deeper questions about whether some technologies should be used at all’ (Katell et al. 2020, 45).

This technical focus can be problematic particularly in its resisting critical reflection of the status quo. Technology-centered reforms, even with an explicit focus on fairness, risk legitimizing structural disadvantage by invoking false objectivity, consequently disincentivizing reforms that would restructure socio-technical systems as opposed to their technical constituents (Green 2020). These dynamics have been discovered in ability-grouping practices which ‘continue to abound in schools and are legitimated with data’ (Datnow and Park 2018, 148) even though their ill effects, primarily affecting those in positions of existing disadvantage, are widely recognized in scholarship. “Algorithmic fairness” may further legitimate inequitable practices through nominal certifications of fairness, as I argue below, but it may also promote technical interventions even though ‘many meaningful interventions toward equitable algorithmic systems are non-technical’ and localized as opposed to scalable (Katell et al. 2020, 45).

4.2. Fairness in conditions of structural inequality

Ideals of meritocracy and formal equality (e.g., equal treatment and equal opportunity) have long roots in the operative meanings of educational accountability (see Fontaine [2016]). Fair ML metrics that condition fairness on actual outcomes in the data reflect these notions (Kizilcec & Lee [forthcoming]; Wachter, Mittelstadt, and Russell 2021), and are attractive from the perspective of high-stakes accountability, respectively. *Calibration*, for example, ensures that teachers and administrators can both trust their algorithm-supported decisions and evince that every learner is subjected to the same evaluation criteria. It cannot usually be satisfied simultaneously with other formal equality metrics, such as *Equalized Odds*, however (Kleinberg, Mullainathan, and Raghavan 2017). Notably, use of protected attributes as model features could perhaps resolve this trade-off, but the use of demographic data is typically found unattractive due to concerns about legal compliance (see Hellman [2020]; Carmel & Ben-Shahar [2017]). Moreover, the use of behavioural data which, in contrast to demographic data, allegedly reflects factors for which students are accountable for is taken to be better aligned with meritocratic ideals (Whitman 2020).

I wish to make no normative arguments regarding these issues here, nor to downplay the related moral and legal tensions. I only note that, in conditions of structural inequality, policies and metrics grounded in ideals of meritocracy and formal equality can work against *substantive equality* which would recognize that students start from different positions of (dis)advantage. U.S. education policy, for instance, has shown how ‘efforts to ignore race via ‘colorblind’ or race-neutrality policies such as school choice or accountability systems can easily replicate rather than address age-old patterns of inequality grounded in a history of race consciousness’ (Wells 2014, i). These tendencies are reflected in fair ML metrics that condition ideal balances in statistical metrics on actual outcomes in the data. *Calibration* and *Equalized Odds*, for example, can be satisfied even when they preserve structural bias, such as “achievement gaps” traceable to segregation and differences in access to study resources between students from low- and high-income families, for example. They deem algorithms ‘fair’ insofar as such gaps are not further widened. Wachter, Mittelstadt, and Russell (2021) call these metrics ‘bias-preserving’ because they take the status quo, existing inequalities included, as a neutral baseline for fairness. Differences between these metrics pertain only to *how* the structural bias is preserved. ‘Bias-transforming’ metrics (e.g., *Statistical Parity*), on the other hand, cannot be satisfied in conditions of structural inequality even by perfectly accurate algorithms (ibid.).

Bias-preserving metrics inherently couple predictive accuracy with fairness and thereby seem a natural choice within accountability systems that prioritize ideals of meritocracy and formal equality. The relative merits of these metrics notwithstanding (see Kizilcec & Lee [forthcoming]), a decision to use bias-preserving metrics is one that effectively takes existing social stratification and inequality across compared groups as justified. Neglecting how forms of (dis)advantage structure educational datasets, bias-preserving metrics construct structural injustice as a causal influence on learners’ prospects that agents in the educational domain have no responsibility to mitigate. Socio-material conditions and structural injustices against historically oppressed groups are ‘disentangled from student performance’ (Fontaine 2016, 8) in the “fair” model. Conversely, bias-transforming metrics with their grounding in substantive equality will indicate unfairness even when predictions match structurally biased historical distributions in data. As substantive equality is required by, for example, EU non-discrimination legislation in certain jurisdictions, these metrics may, notably, be even required for compliance (Wachter, Mittelstadt, and Russell 2021).

5. Concluding remarks

This article has discussed algorithmic fairness as an emerging constituent of (algorithmic) accountability in education, critically examining formal and technological tools for detecting and mitigating algorithmic bias. In the market for algorithmic accountability instruments, the alleged transportability, universality, comparability, and objectivity of fair ML frameworks make them attractive to educational institutions keen on demonstrating the effectiveness and equitable impact of implemented AIED technologies and legitimating their use in the eyes of the public.

I have argued that specific meanings of (in)equality, (non-)discrimination, and (un)fairness are constructed through algorithmic fairness analyses, reflecting, and actively (re)configuring existing meanings and performance of accountability in education. Specifically, three claims were made: First, fairness analyses aim to objectively uncover and disclose the impact algorithms have on equality, but much discretionary ethically and politically significant decision-making precedes, underlies, and follows such analyses. Consequently, transparency regarding mere absence of certain biases does not necessarily disclose meaningful aspects of AIED systems to the public, assuming transparency can even be achieved given private vendors’ discretion. Second, given the formalistic epistemology of fair ML methodologies, even fairness-conscious adoption of AIED systems risks resulting in misguided or insufficient interventions. Lastly, if constrained by, and harnessed for satisfying accountability policies that favour formally equal and meritocratic treatment in conditions of structural injustice, “algorithmic fairness” may disincentivize systemic reforms for substantive equality in

education. As this study has been largely conceptual, future empirical research is needed to further assess these claims.

I emphasize that my claim has not been that algorithmic fairness analyses would be unhelpful or useless: they can surely support risk management, especially when multiple metrics are used (Kizilcec and Lee forthcoming), and they can make the politics of AIED systems more transparent. Rather, I have aimed to critically examine and highlight the political, epistemological, and ethical dimensions and implications of fairness analyses. This study's findings support previous work suggesting that, even with their language of excellence and equity, and an emphasis on informed, data-driven decision-making, certain accountability systems and incentive structures may paradoxically reproduce and legitimize existing inequality. This study has contributed by showing how explicit measures taken to govern and account for AIED technologies themselves may suffer from a similar second-order tendency. Accordingly, as methods, tools, and policies for governing algorithms in education begin to emerge, scholarship is called to critically examine the ethics and politics that underlie them.

Notes

1. However, differential item functioning (DIF) has been largely neglected in fair ML discussions (Hutchinson and Mitchell 2019).
2. <http://aequitas.dssg.io/>.
3. <https://aif360.mybluemix.net/>.
4. Ground truth data, especially false negatives, can often be unavailable, however.

Acknowledgments

This work was supported by Academy of Finland under grant number 326584. I thank Nelli Piattoeva and the anonymous reviewers for their insightful comments and feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Academy of Finland: [Grant Number 326584].

Notes on contributor

Otto Sahlgren is a PhD candidate at Tampere University, working on technology ethics and the philosophical underpinnings of fair machine learning research.

ORCID

Otto Sahlgren  <http://orcid.org/0000-0001-7789-2009>

References

- Adams, Richard, Sally Weale, and Caelainn Barr. 2020. "A-level Results: Almost 40% of Teacher Assessments in England Downgraded". *The Guardian*, August 13th. <https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded>.
- Baker, Ryan S., and Aaron Hawn. 2021. "Algorithmic Bias in Education." EdArXiv, March. doi:10.35542/osf.io/pbmzv.
- Binns, Reuben. 2018. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31 (4): 543–556. doi:10.1007/s13347-017-0263-5.

- Boaler, Jo, Dylan Wiliam, and Margaret Brown. 2000. "Students' Experiences of Ability Grouping-Disaffection, Polarisation and the Construction of Failure." *British Educational Research Journal* 26 (5): 631–648.
- Bovens, Mark. 2007. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13 (4): 447–468.
- Camilli, Gregory. 2013. "Ongoing Issues in Test Fairness." *Educational Research and Evaluation* 19 (2–3): 104–120. doi:10.1080/13803611.2013.767602.
- Capano, Giliberto, and Andrea Lippi. 2017. "How Policy Instruments are Chosen: Patterns of Decision Makers' Choices." *Policy Sciences* 50 (2): 269–293. doi:10.1007/s11077-016-9267-8.
- Carmel, Yoni Har, and Tammy Harel Ben-Shahar. 2017. "Reshaping Ability Grouping Through Big Data." *Vanderbilt Journal of Entertainment & Technology Law* 20 (1): 87–128.
- Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5 (2): 153–163. doi:10.1089/big.2016.0047.
- Coleman, James S. 1975. "What Is Meant by 'An Equal Educational Opportunity'?" *Oxford Review of Education* 1 (1): 27–29.
- Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*: 797–806.
- Datnow, Amanda, and Vicki Park. 2018. "Opening or Closing Doors for Students? Equity and Data use in Schools." *Journal of Educational Change* 19 (2): 131–152. doi:10.1007/s10833-018-9323-6.
- Desrosières, Alain. 2001. "How Real Are Statistics? Four Possible Attitudes." *Social Research* 68 (2): 339–355.
- Diamond, John B., and James P. Spillane. 2004. "High-stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *Teachers College Record* 106 (6): 1145–1176.
- Digital Promise. 2014. *Improving Ed-Tech Purchasing: Identifying the Key Obstacles and Potential Solutions for the Discovery and Acquisition of K-12 Personalized Learning Tools*. <https://digitalpromise.org/2014/11/13/improving-ed-tech-purchasing>.
- Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness". *Proceedings of the 3rd innovations in theoretical computer science conference*: 214–226. doi:10.1145/2090236.2090255.
- Eder, Donna. 1981. "Ability Grouping as a Self-Fulfilling Prophecy: A Micro-Analysis of Teacher-Student Interaction." *Sociology of Education* 54 (3): 151–162. doi:10.2307/2112327.
- Fazelpour, Sina, and Zachary C. Lipton. 2020. "Algorithmic Fairness from a Non-ideal Perspective". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*: 57–63. doi:10.1145/3375627.3375828.
- Felouzis, Georges, and Samuel Charmillot. 2013. "School Tracking and Educational Inequality: A Comparison of 12 Education Systems in Switzerland." *Comparative Education* 49 (2): 181–205. doi:10.1080/03050068.2012.706032.
- Fontaine, Claire. 2016. "The Myth of Accountability: How Data (mis) use is Reinforcing the Problems of Public Education". Data & Society Research Institute. https://www.datasociety.net/pubs/ecl/Accountability_primer_2016.pdf.
- Green, Ben. 2020. "The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness". *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 594–606. doi:10.1145/3351095.3372869.
- Green, Ben. 2021. "The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action". arXiv preprint arXiv:2106.01784.
- Green, Ben, and Lily Hu. 2018. "The myth in the Methodology: Towards a Recontextualization of Fairness in machine learning". *Proceedings of the Machine Learning: The Debates workshop*.
- Hamilton, Laura S., Brian M. Stecher, and Stephen P. Klein, eds. 2002. *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: RAND Education.
- Hanna, Alex, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. "Towards a Critical Race Methodology in Algorithmic Fairness". *Proceedings of the 2020 conference on fairness, accountability, and transparency*: 501–512. doi:10.1145/3351095.3372826.
- Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning". *Proceedings of the 30th International Conference on Neural Information Processing Systems*: 3323–3331.
- Hartong, Sigrid, and Andreas Breiter. 2020. "Between Fairness Optimization and 'Inequalities of Dataveillance': The Emergence and Transformation of Social Indices in German School Monitoring and Management." In *World Yearbook of Education 2021: Accountability and Datafication in the Governance of Education*, edited by Sotiria Grek, Christian Maroy, and Antoni Verger, 54–71. London: Routledge.
- Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106 (4): 811–866.
- Hutchinson, Ben, and Margaret Mitchell. 2019. "50 Years of Test (un) Fairness: Lessons for Machine Learning". *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 49–58. doi:10.1145/3287560.3287600.
- The Institute for Ethical AI in Education. . 2021. *The Ethical Framework for AI in Education*. <https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/>.
- Jarke, Juliane, and Andreas Breiter. 2019. "Editorial: The Datafication of Education." *Learning, Media and Technology* 44 (1): 1–6. doi:10.1080/17439884.2019.1573833.

- Kamiran, Faisal, Asim Karim, and Xiangliang Zhang. 2012. "Decision Theory for Discrimination-aware Classification". 2012 *IEEE 12th International Conference on Data Mining*: 924–929. doi:10.1109/ICDM.2012.45.
- Katell, Michael, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. 2020. "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field". *Proceedings of the 2020 conference on fairness, accountability, and transparency*: 45–55. doi:10.1145/3351095.3372874.
- Kippin, Sean, and Paul Cairney. 2021. "The COVID-19 Exams Fiasco Across the UK: Four Nations and two Windows of Opportunity." *British Politics* 2021: 1–23. doi:10.1057/s41293-021-00162-y.
- Kizilcec, R. F., and H. Forthcoming Lee. "Algorithmic Fairness in Education." In *Ethics in Artificial Intelligence in Education*, edited by Wayne Holmes, and Kaska Porayska-Pomsta. Taylor & Francis.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-offs in the Fair Determination of Risk Scores". *Proceedings of 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. doi:10.4230/LIPIcs.ITCS.2017.43.
- Lascoumes, Pierre, and Patrick Le Galès. 2007. "Introduction: Understanding Public Policy Through its Instruments —from the Nature of Instruments to the Sociology of Public Policy Instrumentation." *Governance* 20 (1): 1–21.
- Levin, Henry M. 1974. "A Conceptual Framework for Accountability in Education." *The School Review* 82 (3): 363–391.
- Luckin, Rose, Wayne Holmes, Mark Griffiths, and Laurie B. Forcier. 2016. *Intelligence Unleashed: An Argument for AI in Education*. London: Pearson Education.
- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54 (6): 1–35. doi:10.1145/3457607.
- Narayanan, Arvind. 2018. "21 Fairness Definitions and Their Politics". Tutorial at the *Conference on Fairness, Accountability, and Transparency*.
- Ozga, Jenny. 2013. "Accountability as a Policy Technology: Accounting for Education Performance in Europe." *International Review of Administrative Sciences* 79 (2): 292–309. doi:10.1177/0020852313477763.
- Ozga, Jenny. 2019. "The Politics of Accountability." *Journal of Educational Change*, 1–17. doi:10.1007/s10833-019-09354-2.
- Piattoeva, Nelli. 2015. "Elastic Numbers: National Examinations Data as a Technology of Government." *Journal of Education Policy* 30 (3): 316–334. doi:10.1080/02680939.2014.937830.
- Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now Institute, April: 1–22. <https://ainowinstitute.org/aiareport2018.pdf>.
- Selbst, Andrew D., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems". *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 59–68. doi:10.1145/3287560.3287598.
- Sellar, Sam. 2015. "Data Infrastructure: A Review of Expanding Accountability Systems and Large-Scale Assessments in Education." *Discourse: Studies in the Cultural Politics of Education* 36 (5): 765–777. doi:10.1080/01596306.2014.931117.
- Verger, Antoni, Clara Fontdevila, and Lluís Parcerisa. 2019. "Reforming Governance Through Policy Instruments: How and to What Extent Standards, Tests and Accountability in Education Spread Worldwide." *Discourse: Studies in the Cultural Politics of Education* 40 (2): 248–270. doi:10.1080/01596306.2019.1569882.
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2021. "Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law." *West Virginia Law Review* 123 (3), URL = https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.
- Wells, Amy S. 2014. *Seeing Past the "Colorblind" Myth: Why Education Policymakers Should Address Racial and Ethnic Inequality and Support Culturally Diverse Schools*. National Education Policy Center. <http://nepc.colorado.edu/publication/seeing-past-the-colorblind-myth>.
- Whitman, Madisson. 2020. "'We Called That a Behavior': The Making of Institutional Data." *Big Data & Society* 7 (1): 1–13. doi:10.1177/2053951720932200.
- Williamson, Ben, and Nelli Piattoeva. 2019. "Objectivity as Standardization in Data-Scientific Education Policy, Technology and Governance." *Learning, Media and Technology* 44 (1): 64–76. doi:10.1080/17439884.2018.1556215.
- Zeide, Elana. 2020. "Robot Teaching, Pedagogy, and Policy." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 789–803. New York, NY: Oxford University Press.