Algorithmic Bias in Education

Ryan S. Baker, University of Pennsylvania, Graduate School of Education, Philadelphia, PA, 19104
Aaron Hawn, University of Pennsylvania, Graduate School of Education, Philadelphia, PA, 19104

Corresponding Author: Ryan S. Baker, ryanshaunbaker@gmail.com

Abstract

In this paper, we review algorithmic bias in education, discussing the causes of that bias and reviewing the empirical literature on the specific ways that algorithmic bias is known to have manifested in education. While other recent work has reviewed mathematical definitions of fairness and expanded algorithmic approaches to reducing bias, our review focuses instead on solidifying the current understanding of the concrete impacts of algorithmic bias in education—which groups are known to be impacted and which stages and agents in the development and deployment of educational algorithms are implicated. We discuss theoretical and formal perspectives on algorithmic bias, connect those perspectives to the machine learning pipeline, and review metrics for assessing bias. Next, we review the evidence around algorithmic bias in education, beginning with the most heavily-studied categories of race/ethnicity, gender, and nationality, and moving to the available evidence of bias for less-studied categories, such as socioeconomic status, disability, and military-connected status. Acknowledging the gaps in what has been studied, we propose a framework for moving from unknown bias to known bias and from fairness to equity. We discuss obstacles to addressing these challenges and propose four areas of effort for mitigating and resolving the problems of algorithmic bias in AIED systems and other educational technology.

Keywords: algorithmic bias, algorithmic fairness, machine learning, artificial intelligence and education

Algorithmic Bias in Education

1. Introduction

Today, algorithms influence our lives in a wide variety of ways. Computer algorithms that make decisions and predictions are often viewed as inherently fair and objective (M. K. Lee, 2018). But in recent years, a competing perspective has emerged -- the perspective that algorithms often encode the biases of their developers or the surrounding society, producing predictions or inferences that are clearly discriminatory towards specific groups. Examples of algorithmic bias cross contexts, from criminal justice (Angwin et al., 2016), to medicine (O'Reilly-Shah et al., 2020), to computer vision (Klare et al., 2012), to hiring (Garcia, 2016). These limitations appear -- and are particularly salient -- for high-stakes decisions such as predicting recidivism (Angwin et al., 2016) or administering anaesthesia (O'Reilly-Shah et al., 2020).

This limitation is found in educational algorithms as well, an increasing number of papers assert (e.g. Bridgeman et al., 2009, 2012; Ocumpaugh et al., 2014; Yudelson et al., 2014; Kai et al., 2017; Hu & Rangwala, 2020; Yu et al., 2020). The problem of bias in educational testing has been documented since the 1960s and anticipated many aspects of the modern literature on algorithmic bias and fairness (see review and discussion in Hutchinson & Mitchell, 2019). In recent years, algorithms have become applied in educational practices at scale for a range of applications, often high-stakes, including dropout prediction (Christie et al., 2019; Milliron et al., 2014), automated essay scoring (Ramineni & Williamson, 2013), graduate admissions (Waters & Miikkulainen, 2014), and knowledge inference (Ritter et al., 2016). Academics have been warning about possible uneven effectiveness and lack of generalizability across populations in educational algorithms for several years (e.g. Bridgeman et al., 2009; Ocumpaugh et al., 2014). This concern became very salient to the general public in the 2020 UK GCSE and A-Level grading controversy (H. Smith, 2020), where a set of formulas was developed by the national qualifications regulator (by hand rather than by an automated algorithm) to assign predicted examination grades based on teacher predictions -- the algorithm assigned poorer grades to students in state-funded schools and better grades (even better than teacher prediction) to students in smaller independent schools.

In this paper, we review algorithmic bias in education, discussing theoretical work on the root causes of algorithmic bias, and reviewing the existing empirical literature on the specific ways that algorithmic bias is known to have manifested in education. In doing so, we distinguish ourselves from more algorithmically-focused reviews of algorithmic bias in education (e.g. Kizilcec & Lee, 2020). There is a great deal of merit to reviewing mathematical definitions of fairness and algorithmic approaches to reducing bias. Our review focuses instead on understanding exactly who appears to be impacted, and the impact played by the context surrounding the algorithms themselves. We focus in particular on biases emerging from how variables are operationalized and which data sets are used. This review is also distinct from broader discussions of how artificially intelligent technologies can be biased (e.g. Holstein & Doroudi, in press), including in the processes of the design of these technologies, focusing on the narrower issue of bias in the algorithms used to assess and make decisions.

Our review starts in section 2 with a discussion of theoretical and formal perspectives on algorithmic bias. Within that, we define algorithmic bias, and connect it to different perspectives on the machine learning pipeline. We review metrics for assessing bias and taxonomies on forms

bias, and consider the essential link between two forms of bias in data collection -- situating key challenges around algorithmic bias not in the algorithms themselves, but in the data input to those algorithms.

In section 3, we review the evidence around algorithmic bias in education. We start by reviewing the evidence around the most heavily-studied categories (race/ethnicity, gender, and nationality). We then review and discuss the evidence for algorithmic bias in less-studied categories. We then discuss the (very large) gaps in what has been studied, with a focus on the America-centric bias of most of the research on algorithmic bias in education.

Finally, in section 4, we propose a framework for moving from unknown bias to known bias to fairness. We discuss the obstacles to addressing algorithmic bias in education, and propose four directions for the field of AIED to move in, that can help to mitigate and resolve the problem of algorithmic bias, for the benefit of the students our community serves.

2. Theoretical and formal perspectives on algorithmic bias

2.1 Defining algorithmic bias

Discussions of algorithmic bias in education have been complicated by overlapping meanings of the term *bias* (Crawford, 2017; Blodgett et al., 2020). For example, a recent survey of the uses of the term *bias* in natural language processing across 146 papers found several areas for clarification in how authors define and write about bias, from a lack of explanation about how exactly systems were biased to confusion as to the eventual harms these forms of bias might cause (Blodgett et al., 2020). While not attempting to reconcile these conflicts, we will briefly discuss some issues of definition for *algorithmic bias* before proposing a more limited working definition we use in our survey.

Algorithmic bias in emerging use

The term *algorithmic bias* has been applied to an array of examples of unfairness in automated systems, only some of which seem to fit statistical or technical definitions of bias. In response, as examples have appeared in popular use, researchers have attempted to describe criteria for *bias*. Mitchell et al. (2021) summarizes "popular media" usage of the term *biased* as describing cases where "a model's predictive performance (however defined) unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class" (p.1). Similarly, Gardner et al. (2019) use *bias* as "inequitable prediction across identity groups" (p. 228). Applying a broader definition, Suresh and Guttag (2020), refer to *biases* as possible sources of harm throughout the machine learning process, including "unintended or potentially harmful" properties of the data that lead to "unwanted or societally unfavorable outcome[s]" (p. 1-2).

Other authors favor the use of *unfair* over *biased*, preserving *bias* for its statistical meaning and using *fair/unfair* for its social/moral implications. Mehrabi et al. (2019) describe *fairness* in the context of algorithmic decision-making as "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics. An unfair algorithm is one whose decisions are skewed towards a particular group of people" (p. 1). Similarly, while acknowledging that algorithmic bias, in its popular sense, refers to socially objectionable "demographic disparities," Barocas et al. (2019) avoid this popular usage of bias altogether, instead using *demographic disparity* and *discrimination* to refer to the negative impacts of applying some models, while maintaining *bias* in its statistical sense as systematic error in either data or model estimates. Friedman and Nissenbaum's (1996) definition foreshadows these

distinctions, claiming that *biased* computer systems "*systematically* and *unfairly discriminate* against individuals or groups of individuals in favor of others [emphasis in original]" (p. 332).

Overall a range of "biases" appear across definitions, from the statistical biases of measurement and error to imbalances in how well a model performs across groups, to systematic skew in results, to disparate impacts and discrimination as model results are interpreted and applied.
As the nature and impacts of algorithmic bias continue to emerge, Blodgett et al.'s (2020) suggestions for clarity in description of bias are timely, particularly in education, where fewer cases of algorithmic bias impacting real-world outcomes have been reported, compared with criminal justice or healthcare. The authors suggest including descriptions of the kinds of system behaviors found to be harmful, how and why the system behaviors are harmful, who is harmed, and the normative reasons for making these judgements.

Research outside of education has put a spotlight on some of the very real harms that can result from algorithmic bias. Such harms have been categorized broadly into *allocative* and *representational* forms (Crawford, 2017; Suresh & Guttag, 2020). Allocative harms result from the withholding of some opportunity or resource from specific groups or the unfair distribution of a good across groups. Some examples include gender and racial bias in ad delivery for housing and employment (Ali et al., 2019; Benner et al., 2019); gender bias in assigning credit limits (Knight, 2019; Telford, 2019); racial bias in sentencing decisions (Angwin et al., 2016), racial bias in identifying patients for additional health care (Obermeyer et al., 2019), and -- in education -- bias in standardized testing and its resulting impact on high stakes admission decisions (Dorans, 2010; Santelices & Wilson, 2010). Representational harms, on the other hand, manifest as the systematic representation of some group in a negative light, or in a lack of positive representation (Crawford, 2017). Multiple forms of representational harm have been uncovered in recent years. Work by Sweeney (2013) identifies representational harms of *denigration* and *stereotyping*, where the word "criminal" was more frequently returned in online ads after searches for black-identifying first names. Kay et al. (2015) describe other representational harms of *recognition* and *under-representation,* with work finding that image search results for male-dominated professions displayed a higher proportion of males than the proportions suggested by the data from U.S. Bureau of Labor and Statistics.

As can be seen, there are a range of ways that algorithmic bias can be defined and viewed. In the current article, we focus on studying algorithmic bias in terms of situations where model performance is substantially better or worse across mutually exclusive groups (i.e. Gardner et al., 2019; Mehrabi et al., 2019; Mitchell et al., 2021). Other forms of algorithmic bias can also be highly problematic, but -- as we discuss below -- the published research in education thus far has focused on this form of algorithmic bias. In this review, we also focus on the bias in algorithms, not in the broader design of the learning or educational systems that use these algorithms. Bias can emerge in the design of learning activities, leading to differential impact for different populations (Finkelstein et al., 2013), but this is a much broader topic than the bias that comes from algorithms.

Bias against whom?

Researchers have also considered a range of groups which have been, or might be, impacted by algorithmic bias. Many of these groups have been defined by characteristics protected by law, in various countries. In the United Kingdom, for instance, the Equality Act 2010 merged over a hundred disparate pieces of legislation into a single legal framework, unifying protections

against discrimination on the basis of sex, race, ethnicity, disability, religion, age, national origin, sexual orientation, and gender identity. In the United States, the same categories are protected by a combination of different legislation, commission rulings, and court rulings, dating back to the Civil Rights Act of 1964. Similar laws afford protections in the European Union and most other countries around the world, though differing somewhat in which groups are protected and how groups are defined.

However, while critical for social equity, looking for bias only under the lamppost of nationally-protected classes (categories often with their own complicated histories) may be leaving unexamined serious impacts on other, under-investigated, groups of people. Depending on the social context of specific algorithmic systems, other researchers have suggested additional characteristics vulnerable to bias, such as urbanicity (Ocumpaugh et al., 2014), military-connected status (Baker et al., 2020), or speed of learning (Doroudi & Brunskill, 2019). Section 3.2 reviews the limited education research into algorithmic bias associated with other groups.

While recognizing that many of the implications and harms of algorithmic bias currently fall outside of legal purview, legal frameworks used to decide which classes of people merit protection from discrimination may be helpful in assessing the unknown risks that algorithmic bias poses to as yet unidentified groups (Soundarajan & Clausen, 2018). Past discriminatory patterns against individuals, as well as the degree that class membership can be changed, may be useful in focusing efforts in bias mitigation. Judicial standards may also be helpful, particularly in medicine and education, in determining whether or not discrimination is justified by a compelling interest. Machine-learning applications that diagnose medical conditions or generate Early Warning Indicators (EWIs) for high school dropout could potentially satisfy criteria justifying discrimination when the benefits of an application outweigh possible harms, when individuals are aware of algorithmic decisions and are able to opt out or modify them, and when interventions are narrowly tailored to the relevant features of a targeted class (Soundarajan & Clausen, 2018).

A note on the term "algorithmic"

In this review, most examples of algorithmic bias are drawn from cases where large datasets of past examples were used to train predictive models using algorithmic or statistical methods. In other work, the term *algorithmic* in this context has been used to mean automated decision-making systems in general, with machine-learning products and processes playing an increasing role in such automated systems. While this article uses *algorithmic* to refer more specifically to processes where models were created by applying automated algorithms to data, the recent A-Level grading controversies in the UK make clear that bias and discrimination can result in cases where models and automation were coded by hand as well. While most of the cases that we review involve automated algorithms rather than manually-created models, many of the themes we discuss are relevant to manually-created models as well.

2.2 Origins of bias and harm in the machine learning pipeline

Researchers in education have recently begun drawing attention to existing cases of algorithmic bias in educational technologies, as well as to the possible increase, through adoption of artificial intelligence and machine learning, of such bias and its harmful impacts. As a starting place in understanding the origins of algorithmic bias, Mitchell et al. (2021) make a helpful distinction between *statistical* and *societal* forms of bias, where *statistical bias* encompasses sampling bias

and measurement error and *societal bias* refers to "concerns about objectionable social structures that are represented in the data" (p. 4). In a model used in practice, either of these forms of bias might contribute to overall algorithmic bias and possible real-world discrimination and harms. Several authors working at the intersection of computer science and the social sciences have laid the groundwork for this examination, expanding the statistical/societal distinction to describe the kinds of bias that can arise at each stage of the machine learning lifecycle and categorizing the possible harms and impacts arising from that bias.

In the context of fairness and bias, process descriptions of machine learning range from the simple to the complex. Some collapse the process into broad stages, such as Barocas et al. (2019) and Kizilcec and Lee (2020), who view the process as going from measurement to model learning to action, or Mehrabi et al. (2019), who considers the process as going from user interaction to data to algorithm. Others, such as Suresh and Guttag (2020), have conceptualized machine learning in a finer-grained way and have considered the possibility of bias at each of the following stages: *data collection, data preparation, model development, model evaluation, model post-processing*, and *model deployment*.

There have also been efforts in collaboration between industry and academia to develop process maps to support decision-making around bias in organizational contexts. One such process, illustrated in Figure 1, includes additional stages for *Task Definition* prior to *Dataset Construction*, along with separate stages for a *Testing Process* prior to *Deployment*, and ongoing *Feedback* from users (Cramer et al., 2019). The authors contend that explicitly considering each of these stages is critical in an organizational response to mitigating bias.

Common sources of bias have been outlined in several taxonomies and related to stages in the machine learning lifecycle (Barocas et al., 2019; Hellström et al., 2020; Mehrabi et al., 2019; Silva & Kenney, 2018; Suresh & Guttag, 2020). Landmark early work by Friedman and Nissenbaum (1996) set a precedent for this sequential framing of bias, categorizing biases based on the context where they arise, whether from preexisting social biases, technical limitations, or use of the models by real users. As this early work makes clear, much of the bias ultimately detected in algorithms arises outside the actual training of the model, whether from preexisting historical bias, aspects of measurement and data collection, or the uses to which model predictions are put.

More exhaustively, Suresh and Guttag (2020) categorize bias into *Historical bias, Representation bias, Measurement bias, Aggregation bias, Evaluation bias*, and *Deployment bias* and map these forms of bias to stages in the machine learning process (See Figure 1 for a complete mapping). *Historical bias* involves mismatches between the world as it is and values for how the world should be, causing models that replicate decisions made in the world to be biased. An example in education might be including student demographics as predictors of grades and generating a model that achieves better performance by using those features to predict lower grades for some students -- i.e. if a student is in group X, they get 5 points lower on their final grade (i.e. Wolff et al., 2013). Such a model achieves better accuracy at the cost of potentially perpetuating bias. Bias of this nature can also arise even when demographics are not explicitly encoded in models, for instance when a proxy for a demographic variable is unintentionally included as a predictor. This form of bias is surprisingly common in education -- in a recent survey of the role of demographics in the Educational Data Mining (EDM) community, Paquette et al. (2020) found that roughly half of papers including demographics in

analyses used at least one demographic attribute as a predictive feature within the model, without incorporating demographics into model testing or validation.

Representational bias is when a group that is under-sampled in the training data is predicted more poorly; for example, Anderson et al. (2019) find that the college graduation prediction they develop works poorly for indigenous learners, who comprise a very small proportion of the overall sample. *Measurement bias* occurs when choosing variables, for example when variables do not have construct validity for what they are intended to represent, and biases in the variables chosen cause unequal prediction across groups. For example, a model predicting school violence may be biased if the labels of which students engage in violence are created by a process that involves prejudice – e.g. the same violent behavior by two students is documented and punished for members of one race but not for members of another race (Bireda, 2002). As Holstein & Doroudi (in press) note, disparities in access to learning technologies can in turn create representational biases, lowering the quality of learning technologies for students who already have less access to them.

Moving past the data collection and preparation phases of machine learning and into model development, Suresh and Guttag (2020) discuss how *aggregation bias* occurs when distinct populations are combined in the same model and the resulting model does not work for some -- or all -- groups of learners. This is seen in Ocumpaugh et al. (2014), where detectors of student emotion trained on a combination of urban, rural, and suburban students function more poorly for all three groups than detectors trained on individual groups. *Evaluation bias* occurs when the test sets used to evaluate a model do not represent the eventual population where the model will be applied. As reviewed in Paquette et al. (2020), many models in educational data mining are developed on non-representative populations, and many papers do not even report what populations the models were tested on, making detection of evaluation bias quite difficult. Finally, *deployment bias* involves a model being used in inappropriate ways -- being designed for one purpose and then used for a different purpose, such as using a model designed to identify student disengagement for formative purposes to assign participation grades.

While raising overlapping concerns with Suresh and Guttag (2020), Olteanu et al. (2019) highlight the potential for bias specifically in the analysis of social data, such as user-generated content, online behavior and networks. To the degree that education algorithms leverage data from social software functions, they may be vulnerable to some of the same biases as analysis of social media in other contexts. The authors frame bias in a statistical sense, as a threat to the validity of research conclusions, where proof/disproof of hypotheses is undermined by biases that threaten the internal, external, or construct validity of the research findings. However, they note how examples of statistical bias may relate to the systemic, discriminatory biases which are often thought of in connection to algorithmic bias.

Bias of the forms discussed above has the potential to manifest in a range of educational applications. Algorithmic bias has been documented in situations ranging from at-risk prediction for high school or college dropout (Anderson et al., 2019), at-risk prediction for failing a course (Hu & Rangwala, 2020; H. Lee & Kizilcec, 2020), automated essay scoring (Bridgeman et al., 2009, 2012), assessment of spoken language proficiency (Wang et al., 2018), and even the detection of student emotion (Ocumpaugh et al., 2014). There has been documentation of algorithmic bias impacting educational algorithms in terms of student race, ethnicity, nationality, gender, native language, urbanicity, parental educational background, socioeconomic status, and

whether a student has a parent in the military. We review the literature on these specific findings in Section 3 below.

Figure 1 summarizes several of the machine learning life-cycles mentioned previously alongside sources of bias that might arise at each stage. Where possible the broad categories of measurement, learning, and action on the vertical axis align with taxonomies from Barocas et al. (2019) and Kizilcec and Lee (2020). Sources of bias are drawn from multiple papers (Cramer et al., 2019; Mehrabi et al., 2019; Olteanu et al., 2019; Mitchell et al., 2021; Paullada et al., 2020; Suresh and Guttag, 2020), indicated by a bracketed number within the figure.

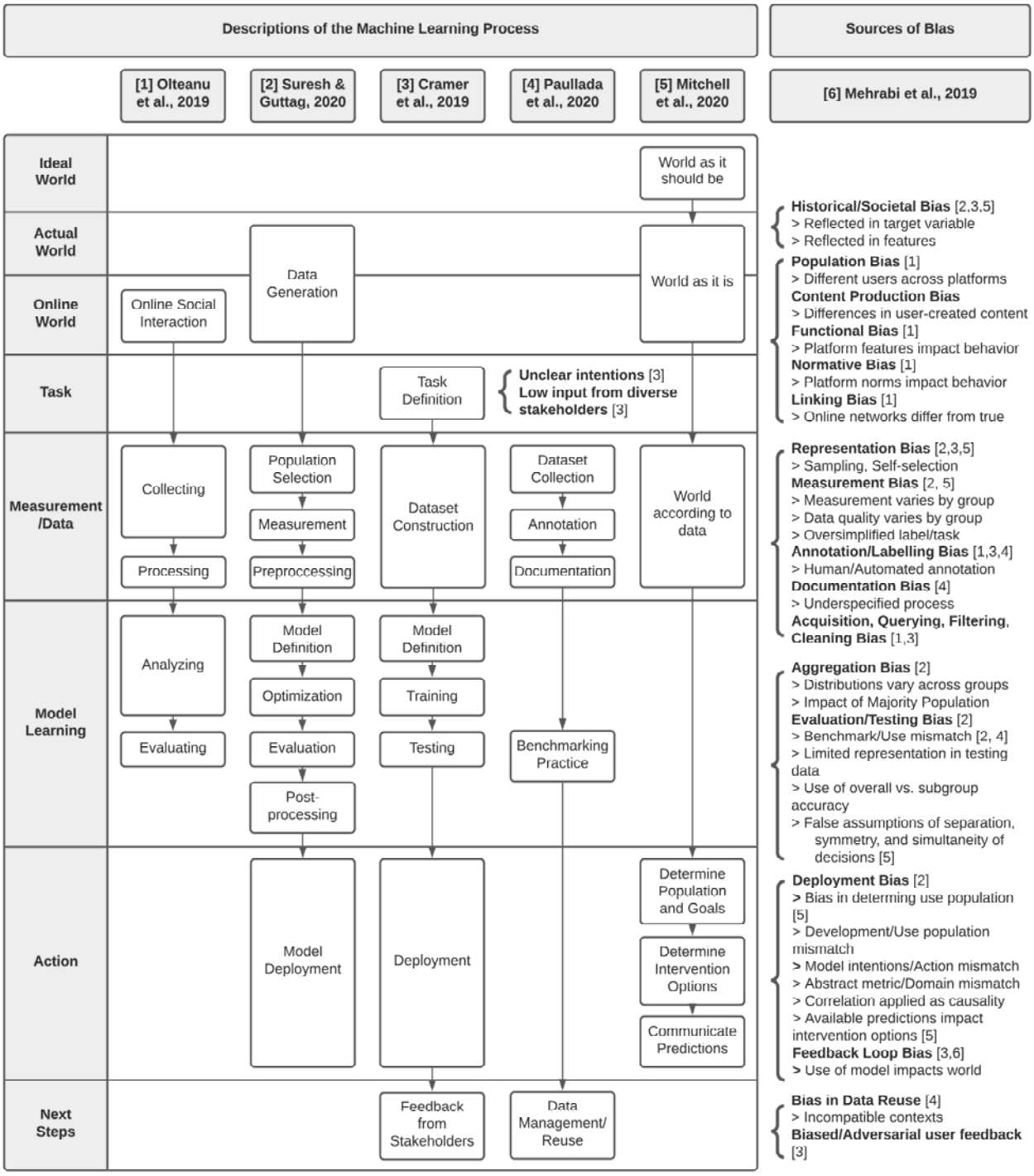

Figure 1: Descriptions of the Machine Learning Process and Possible Sources of Bias

Alternative schema for the origins of bias

While several attempts have been made at locating sources of bias within the machine learning pipeline (See Figure 1), other researchers have argued for locating algorithmic bias, not only at a stage within this process, but also as the product of the social interactions surrounding the production and use of an algorithm. Drawing from sociocultural activity theory, Ferrero and

Barujel (2019) describe an algorithm as the artifact of an activity system, locating bias within the connected parts of that system, where subjects make decisions within a context of objects, rules, community, and division of labor. Applying this alternative framework, they identify biases as either *theoretical, methodological, interpretive, due to decontextualization*, or *due to data training*. The authors' description of bias as the artifact of specific decisions by particular agents draws attention to the fact that identifying a stage at which bias is generated goes only part of the way towards mitigating that bias. The activity system view elaborates other possible pathways for eliminating bias by grounding the temporal stages of a process model in the decisions of subjects working with organizational constraints. Ferrero and Barujel's (2019) activity framing also encourages an examination of algorithmic bias as it develops across contexts. As commonly happens in education, one organization, perhaps commercial, develops a predictive algorithm, while another, a school or district, applies it (see discussion of this issue also in Holstein & Doroudi, in press). Locating machine learning in a similarly broad context, Dieterle et al. (under review) describe algorithmic bias as one in a larger series of AI-driven *divides* in educational technology. The authors describe how a digital divide in access to online learning leads to divides in the representativeness of data across populations, in how algorithms are developed, in how data are interpreted in schools, and eventually in how civic society is impacted.

The increasing quantity of research and journalism over the last decade, empirically describing cases of algorithmic bias and their far-reaching harms, has galvanized public awareness and prompted extensive academic and industry research into the ways that algorithmic bias can be more effectively identified, mitigated, and its harms reduced.

2.3 Formal fairness and its application (in a messy world)

Much of this recent work addressing algorithmic bias has focused on mitigating bias at the model evaluation and postprocessing stages of the machine learning pipeline. More specifically, many articles have aimed at developing and cataloging a variety of formalized fairness definitions and metrics against which models might be evaluated. Recent surveys, both in education and more broadly, present different taxonomies of fairness and its measurement (Barocas et al., 2019; Caton & Haas, 2020; Kizilcec & Lee, 2020; Mehrabi et al., 2019; Mitchell et al., 2021; Verma & Rubin, 2018).

Several researchers (Barocas et al., 2019; Kizilcec & Lee, 2020; Verma & Rubin, 2018) divide mathematical formalizations of fairness into high-level categories of statistical, similarity-based, and causal definitions. Statistical measures examine fairness through the lens of the confusion matrix, calculating relationships and ratios between predicted or true, positive or negative values in relation to membership in a sensitive group. Many of these formalized criteria for fairness connect back to three possible relationships of conditional probability across three main variables: the prediction, the true outcome, and membership in the relevant group. For example, the criteria of *independence,* requires simply that the algorithm's decision or score be independent of membership in the group under consideration, with specific measures including statistical parity, group fairness, and demographic parity (Barocas et al., 2019; Kizilcec & Lee, 2020). Judged against this criterion, an independent (and therefore fair) algorithm might predict that equal percentages of students belonging to different demographic groups will fail a college course or that urban and rural students will have equal knowledge of farming technologies. *Independence* can be useful in some contexts, particularly when there are laws or regulations requiring that individuals be admitted in equal proportion of racial or gender categories (Chouldechova, 2017;

Makhlouf et al., 2020). However, this criterion has its limitations in not considering true outcomes in relation to predicted values; depending on context, this definition of fairness may correct for the historical biases impacting real-world data, or may produce undesirable results until those historical biases have been addressed.

Two other criteria of statistical fairness, *separation* and *sufficiency*, consider group membership and algorithmic predictions, while also incorporating the true outcome of the predicted variable. Generally, *separation* requires that an algorithm's performance be fair across groups, or stated another way, that correct and incorrect predictions are distributed equally in relation to the groups under consideration. This criterion has generated several attempts at precise formulation, from equal opportunity/equalized odds (Hardt et al., 2016) to slicing analysis (Gardner et al., 2019) to predictive parity (Chouldechova, 2017). The *sufficiency* criterion asks the slightly different question of whether true outcomes are independent of group membership, while taking into account the predicted values. Satisfying s*ufficiency* would require, for example, that of all students predicted to drop out in some educational setting, the same proportion are correctly predicted across sensitive groups (Kizilcec & Lee, 2020).

In addition to these statistical definitions, other high-level categories of fairness include similarity-based measures, which account for similarities between individuals beyond a single sensitive attribute (Dwork et al., 2012; Verma & Rubin, 2018), and causal measures, which apply causal graphs and structural equation modeling to trace relevant pathways between a sensitive group attribute through intermediate factors to the predicted outcome (Verma & Rubin, 2018).

Along with this expanded set of formalized metrics and their clarifications of algorithmic bias has come the recognition that applying fairness measures in practice reveals its own range of obstacles. Specifically, technical obstacles to the use of fairness metrics manifest in several "impossibility" results (Chouldechova, 2017; Kleinberg et al., 2017; Berk et al., 2018; Loukina et al., 2019; Lee & Kizilcec, 2020, Darlington, 1971), where satisfaction of one statistical criterion of fairness makes "impossible" satisfaction of another. Kleinberg at al. (2017), for example, describe how the fairness criteria of calibration, balance for the positive class, and balance in the negative class cannot be simultaneously satisfied, except in special cases of perfect prediction and equal base rates, where both groups have an equal proportion of members in the positive class. *Calibration* here refers to the criterion that the predicted probability of a group to achieve the target variable should match the overall proportion of actual positive instances in the same group. Chouldechova (2017) comes to similar conclusions when analyzing the COMPAS dataset (Angwin et al., 2016) in light of competing fairness perspectives of calibration, predictive parity, and the balance of false predictions across groups.

Education-specific analyses have pointed out similar trade-offs in automated scoring for language proficiency exams (Loukina et al., 2019) and predictions of above-median grades in required college courses (Lee & Kizilcec, 2020). As Kleinberg et al. (2017) point out, varied rules for fairness provide slightly different answers to the same general question: are an algorithm's predictions equally effective across groups? Nonetheless, when the presence of the target variable is imbalanced across groups, attempts to satisfy criteria of calibration or predictive parity result in an imbalance in the rate of false positives or false negatives, leading to disparate impacts across groups (Chouldechova, 2017).

While work in this area has clearly acknowledged the need for tradeoffs between fairness metrics, there are fewer attempts to describe optimal tradeoffs in fairness for domain-specific problems (Lee & Kizilcec, 2020; Makhlouf et al., 2020; Suresh & Guttag, 2020). Providing such guidance is clearly difficult, as within even a single task or domain, the goals and uses to which algorithms are assigned can still vary considerably across local contexts, along with the perceptions of fairness among stakeholders.

Similarity-based and causal metrics may face obstacles as well, particularly in identifying sufficient numbers of cases of individuals close enough to each other on selected distance metrics or who vary along the necessary characteristics (Kizilcec & Lee, 2020; Verma & Rubin, 2018). Both similarity-based and causal measures of fairness also depend on either a distance metric or a causal model, both of which may face their own issues of validity and fairness (Kizilcec & Lee, 2020).

In historical perspective, Hutchinson and Mitchell (2019) point out that many formalized definitions of fairness have precedents from the 1960s and 1970s as researchers worked to minimize bias in standardized testing. By the mid-1970s, however, these same researchers recognized persistent confusion in the use of fairness metrics in practice, both among researchers and in communication to the general public (Cole & Zieky, 2001). Questions were raised about applying different fairness metrics in contexts of competing goals and values (Hunter & Schmidt, 1976). Other work noted that formal definitions often disagreed, that it was challenging to find a principled way to select between metrics, and broached the possibility that efforts towards the formalization of fairness might ultimately distract from more direct efforts to address societal problems of equity and justice (Petersen & Novick, 1976).

These critiques of the 1970s foreshadow current sociotechnical critiques that heavy reliance on statistical definitions of fairness will ultimately impede efforts for the just deployment of algorithms in high stakes, real world situations. Work by Green and colleagues (Green, 2020; Green & Hu, 2018; Green & Viljoen, 2020), for example, suggests that the conflation of fairness defined as statistical measures of group parity with fairness in society creates its own obstacles to addressing social inequities. A focus on statistical solutions, they argue, offers decision-makers a seemingly objective criterion on which to evaluate fairness. Such a criterion could lead developers and users of algorithms to avoid grappling with other consequences of employing algorithms for high-stakes decisions. Such an over-dependence on fairness metrics, Green and colleagues contend, could increase the possibility for additional forms of societal unfairness, such as the limiting or reweighting of the criteria that stakeholders use for high stakes decisions, perpetuating harm through a reliance on historically-biased data, or failing to consider a sensitive group's needs independent of a balance of fairness with other groups.

As these authors suggest, it is important for data scientists and AI researchers to push the boundaries of investigation beyond a canonical understanding of algorithms to an understanding that includes algorithmic interventions and application in context (Green & Viljoen, 2020). Towards that end, it may also be highly important to also focus on identifying and mitigating bias in the earlier stages of the machine learning lifecycle, the stages of data collection and data preparation. There has been extensive ML research addressing the later stages of the ML pipeline: model development, model evaluation, model postprocessing. There may also be considerable gains towards fairness to be found by examining and mitigating bias from the upstream portions of the cycle.

2.4 Representational and measurement biases: the key role for data collection

Much of the work studying algorithmic bias, both in education and beyond, has focused on how algorithms and metrics can be used to assess and remove algorithmic bias (see review in Kizilcec & Lee, 2020). Clearly, there is an important role for both of these types of work. Finding better metrics to assess algorithmic bias -- and pushing the field to use them -- can help to catch many forms of algorithmic bias. Algorithms that attempt to ensure fairness, or at least reduce the likelihood or degree of bias, can play an important role in improving outcomes for students who might otherwise be disadvantaged.

However, attempts to address biases through adjustments to algorithms -- even if the biases are identified at this point -- may be ineffective if we have not collected the right data. Both representational and measurement bias, to use the terminology from Suresh and Guttag (2020), can prevent methods further down the pipeline from being able to detect or resolve bias.

In terms of representational bias, if we do not collect data from the right sample of learners, we cannot expect our models to work on all learners. For example, if we collect training data only from suburban upper middle-class children, we should not expect our model to work for urban lower-income students (we may get lucky -- the model may indeed function effectively -- but we lack a basis for believing this).

A simple proportional data set may be insufficient -- there may be groups of learners for whom insufficient data has been collected to develop and validate a model. Even when we collect a seemingly complete data set -- every learner in a university we are developing a learning analytics model for -- this may still be insufficient. For example, Anderson et al. (2019) noted that their sample of all undergraduates in their university only had 44 learners of indigenous descent, a sample that was too small for their models to work reliably. In cases like this one, we may want to group data from multiple universities together so that there are enough learners of a group we want to ensure our models work for. In general, we may need to over-sample (in terms of actual data collection rather than the machine learning technique) learners from less common groups to be confident that we have enough data for each group of learners for whom the model may function less effectively.

And, all too often, we are unable to sample proportionally, particularly when sampling across institutions. As Baker (2019) notes, it is often much harder to collect data in some schools than others. Some urban school districts have extensive protections and limitations on research that make it far more costly and difficult to conduct research or collect data involving their students. Some rural school districts are sufficiently remote to make it highly costly and difficult to conduct studies in person. Some schools and even universities are too poor (or bureaucratic) to have high-quality data systems. Some organizations are simply more welcoming to researchers than others.

Measurement bias can also be a significant challenge that improved metrics or algorithms cannot entirely address. Although Suresh and Guttag (2020) refer to measurement biases in both training labels and predictor variables, the most concerning measurement biases involve training labels. If a training label is biased for some specific population -- for instance, if Black students are more likely to be labeled as engaging in school violence than White students, even for the same behavior -- then it is difficult to determine whether an algorithm works equally well for both groups, or indeed to find any way to be confident that the algorithm's functioning is not biased. The training label's bias may even come from the student themself, if -- for instance -- we are

predicting a self-report variable that is prone to biases in student responses due to factors such as confidence, cultural interpretation, or stereotype threat (Tempelaar et al., 2020). In these cases, finding an alternate variable to predict -- one not as impacted by bias -- may be the best alternative.

In other cases, the measurement bias may emerge when human coders label data that has already been collected, a step in the pipeline which it is easier to exert control over. Sociocultural factors may impact the reliability of human labeling, for instance -- Okur and colleagues (2018) find that American coders labeling the emotions of Turkish students from facial expressions have systematic biases in their interpretation. These problems can go beyond just inaccuracy -- for example, racial biases can enter into value judgments (Kraiger & Ford, 1985).

In the case where a predictor is biased, it may end up serving as a proxy for a group label, in which case it may be best to discard it from consideration. Alternatively, if it seems to have predictive power after controlling for group, it may be possible to create a distilled variable that contains this predictor's variance after partialing out the group variable.

Ultimately, when possible, the best path to addressing both representational and measurement bias is to collect better data -- data that includes sufficient proportions of all groups of interest, and where key variables are not themselves biased. This step is recognized as essential among learning analytics practitioners (Holstein et al., 2019) but is less emphasized among researchers thus far (see, for instance, Paquette et al., 2020, which reports that many papers involving educational algorithms do not even report aggregate demographic information). Holstein and colleagues (2019) argue that, rather than conducting research on less biased algorithms, one of the most important steps for enhancing fairness in the use of algorithms in education would be for researchers to find ways to support practitioners in "collecting and curating" higher-quality data sets. But doing so depends on knowing what groups we need to make sure are represented in the data sets we use to develop models, the focus of our next section.

3. Population Factors and Student Modeling

3.1 What have we learned about algorithmic bias in education from looking at the most commonly studied demographic categories?

The majority of research on actual algorithmic bias in education (as opposed to theoretical algorithmic bias in education) has looked at three categories: race/ethnicity, nationality (comparing learners' current national locations), and gender. In this section, we review this literature and investigate what the field has learned from these investigations. Table 1 lists the studies included in this section and the demographic categories they each address.

Race/Ethnicity

Kai et al. (2017) differentiated performance between African-American and White students in a model predicting student retention in an online college program. They found that a JRip decision tree model performed much more equitably than a J48 decision tree model. Hu and Rangwala (2020) investigated a range of algorithms for predicting if a student is at risk of failing a course, finding that their models generally perform worse for African-American students, but that this result is inconsistent across university courses. Anderson et al. (2019) differentiated performance between students in different racial/ethnic groups in a model predicting six-year college graduation, across five different algorithms. They found that the algorithms generally had higher

false positive rates for White students and higher false negative rates for Latino students. In contrast to these results, Christie and colleagues (2019) found only very minor differences in the quality of a school dropout model by race, possibly because they included proxies for race in their predictors. Lee and Kizilcec (2020) compare an unmodified algorithm to an equity-corrected algorithm, finding that the unmodified algorithm for predicting course grade performs worse for students in underrepresented racial and ethnic groups than for White and Asian students, but that their correction improves several indicators of fairness. Yu and colleagues (2020) studied prediction of undergraduate course grades and average GPA and found that if race data was included in models, students of several racial backgrounds were inaccurately predicted to perform worse than other students, but that this effect dissipated if only clickstream and survey data were included. Yu and colleagues (2021) studied prediction of college dropout, finding that their models have worse true negative rates and better recall for students who are not White or Asian, and also worse accuracy if the student is studying in person. In the case of automated essay scoring, the E-Rater system was reported to inaccurately give 11th grade Hispanic and Asian-American students significantly higher scores than human essay raters, while being more accurate for White and African American students (Bridgeman et al., 2009). This effect did not replicate with followup studies of GRE students using a later version of E-Rater; instead, for some types of essays, E-Rater gave African American students substantially lower scores than human raters did (Bridgeman et al., 2012; Ramineni & Williamson, 2018).

Despite the strong historical underrepresentation of indigenous learners worldwide (James et al., 2008), very little attention has been paid to indigenous learners across these studies. Across these studies, only Anderson et al. (2019) included indigenous learners as one of the categories studied for algorithmic bias; with only 44 indigenous learners in a sample of over 14,000 learners, model performance was very unstable for this group of learners. This omission likely reflects the overall non-inclusion of indigenous learners in the contexts studied. While this does not necessarily indicate intentional omission on the part of the researchers, it indicates a systematic bias towards conducting this research in contexts where indigenous learners are underrepresented.

Nationality (current national location)

In the case of automated essay scoring, the E-Rater system was reported to inaccurately give Chinese and Korean students significantly higher scores than human essay raters on a test of foreign language proficiency, while being more accurate for students of other nationalities (Bridgeman et al., 2009). A replication involving a later version of E-Rater was again found to give Chinese students higher scores than human essay raters (Bridgeman et al., 2012). This study also found that speakers of Arabic and Hindi were given lower scores (Bridgeman et al., 2012). E-Rater was also reported to correlate more poorly and bias upwards in terms of GRE essay scores for Chinese students, despite correlating comparably across 14 other countries (Bridgeman et al., 2009). The SpeechRater system for evaluating communicative competence in English was found to have substantial differences in accuracy for students of different nationalities, with performance particularly low for native speakers of German and Telugu. SpeechRater's evaluations were also found to systematically bias upwards for Chinese students and downwards for German students (Wang et al., 2018).

Ogan and colleagues (2015) built models predicting student learning gains from a mixture of their behaviors related to help-seeking. Models built using data from learners in the Philippines, Costa Rica, and the United States were each more accurate on students from their own countries than for students from other countries.

Li and colleagues (2021) used a large nationally representative dataset collected in 65 countries to predict student achievement on a standardized examination from variables related to student background. They found that a model trained on data from the United States was highly accurate for students from other economically developed countries but less accurate for students from less economically developed countries.

Gender

Kai et al. (2017) differentiated performance between male and female students in a model predicting student retention in an online college program. They found that performance was very good for both groups, and that a JRip decision tree model performed more equitably than a J48 decision tree model, but that the JRip model still had moderately better performance for female students than male students. Hu and Rangwala (2020) investigated a range of algorithms for predicting if a student is at risk of failing a course, finding that their models generally perform worse for male students, but that this result is inconsistent across university courses. Anderson et al. (2019) differentiated performance between male and female students in a model predicting six-year college graduation, across five different algorithms. They found that the algorithms generally had higher false negative rates for male students. Gardner and colleagues (2019) studied MOOC dropout prediction and found that several algorithms studied performed worse for female students than male students. Curiously, they found that this pattern was attenuated for courses with 50-80% male students (but -- again unexpectedly -- worse when there were fewer than 45% male students). Riazy et al. (2020) investigate whether course outcome prediction is impacted by whether a student is male or female. The differences in prediction quality found were very small -- on the order of a percentage point -- and the differences in overall proportion of predicted pass between groups were generally also fairly small and were inconsistent in direction between algorithms. Similarly, Christie and colleagues (2019) found only very minor differences in the quality of a school dropout model by gender. H. Lee and Kizilcec (2020) compare an unmodified algorithm to an equity-corrected algorithm, finding that the unmodified algorithm for predicting course grade performs worse for male students than for female students, but that their correction improves several indicators of fairness. Yu and colleagues (2020) studied prediction of undergraduate course grades and average GPA and found that female students were generally inaccurately predicted to perform better than male students. Yu and colleagues (2021) studied prediction of college dropout, finding that their models have somewhat worse true negative rates for male students, but somewhat better recall for male students taking courses in-person, regardless of whether protected attributes are included in the models. In the case of automated essay scoring, the E-Rater system was reported to be comparably accurate for male and female students for both 11th grade essays and foreign-language proficiency examinations (Bridgeman et al., 2009). This lack of difference replicated in a second set of studies on a later version of E-Rater (Bridgeman et al., 2012).

Notably, despite the relatively large number of studies differentiating male and female students, no study that we are aware of has explicitly looked at algorithmic bias in terms of non-binary or transgender learners (or any category within the space of LBGTQ identities, for that matter).

Table 1 Studies of algorithmic bias involving the most commonly studied demographic categories

	Gender	Ethnicity					Nationality
	Female/ Male	African-American	White	Latino/ Hispanic	Asian	Indigenous Groups	Various
Bridgeman et al., 2009	X	X	X	X	X		X
Bridgeman et al., 2012	X	X					X
Ogan et al., 2015							X
Kai et al., 2017	X	X					
Ramineni & Williamson, 2018		X					
Wang et al., 2018							X
Anderson et al., 2019	X		X	X		X	
Christie et al., 2019	X	X					
Gardner et al., 2019	X						
Hu & Rangwala 2020	X	X					
Lee & Kizilcec, 2020	X	X*	X	X*	X	X*	
Riazy et al., 2020	X						
Yu et al., 2020	X	X	X	X			
Li et al., 2021							X

* These categories were grouped together as under-represented minority students.

3.2 The limited research on other "populations" or "subpopulations"

In the previous section, we investigated the research on three relatively more widely-studied demographic categories: race, current national location, and male/female. These three categories are likely selected due to a combination of being relatively easy to study, and a perception of their importance. These three categories are clearly important, and the research that has been conducted has established the importance of attending to algorithmic bias in education more generally. While we do not wish to dispute or diminish the importance of these categories in any way, we also would caution that an over-dependence on these categories -- or on traditionally-measured categories in general -- may miss algorithmic bias in other categories. In this section, we discuss the more limited research that has been seen into algorithmic bias for other categories of learners.

In doing so, we must establish some sort of selection criterion for which differences should be treated as categories for which algorithmic bias may occur. Clearly, there may be differences in detection accuracy that are not evidence of bias, per-se. For instance, perhaps detectors of student knowledge are less accurate for students who "game the system", intentionally misusing the system to obtain correct answers without learning (e.g. Johns & Woolf, 2006; Wu et al., 2017). This difference in accuracy is not bias, per-se, but a result of the intentional choice of the learner to defeat the model.

We propose a *non-malleability* test for whether a category should be considered in terms of algorithmic bias. A student can choose whether or not to game the system -- and we may be able to design learning experiences that are less often gamed (Baker et al., 2006; Xia et al., 2020). However, the designer of a learning system or an instructor cannot change or influence a student's race, national location, or gender. Similarly, Soundarajan and Clausen (2018) describe how the legal definition of a protected class rests in part on its "immutability," the degree to which an individual can move in and out of the suggested class.

This test suggests other categories that may be relevant: where a student lives (in a finer-grained fashion than national location); national origin; socioeconomic status; native language; disabilities; age; parental educational background; and parent work that affects student mobility (i.e. migrant work or high-risk jobs such as in the military). In the following section, we investigate the research into some of these categories. Table 2 lists the studies included in this section and the populations they address.

Native language and dialect

Differences in model reliability based on a learner's native language have been studied in the context of educational applications of natural language processing, particularly automated essay scoring. Naismith and colleagues (2018) find that a common measure of lexical sophistication is often developed on word lists that are more appropriate for native speakers than second-language learners; while the method is effective at differentiating between learners at different levels of proficiency, there are often systematic differences in ratings between learners in different countries (i.e. Arabic-speaking learners are often rated lower than Chinese-speaking learners with comparable language proficiency). They also found evidence that biases differ between the corpora used by different testing organizations. Interestingly, Loukina and colleagues (2019) find that when conducting automated essay scoring among essays written by individuals from six different countries, training nation-specific models actually leads to different skews between groups, increasing algorithmic bias compared to training on all groups together.

However, we were unable to identify any work on algorithmic bias in terms of learner dialect, although there has been work to develop learning systems that are appropriate for speakers of non-traditional dialects (e.g. Finkelstein et al., 2013). This is in contrast to the relatively greater attention to learner dialect in research on algorithmic bias in domains other than education (Benzeghiba et al., 2007; Blodgett & O'Connor, 2017; Tatman, 2017). It is possible – even likely -- that some of the differences in the performance of essay scoring algorithms for different racial groups discussed above (i.e. Bridgeman et al., 2012; Ramineni & Williamson, 2018) is due to dialectical differences, but this possibility has not yet been systematically investigated in the published literature.

Disabilities

There has been increasing concern about algorithmic bias negatively impacting individuals with a range of disabilities -- including concerns about the failure of AI systems to adequately recognize gestures of people with mobility or posture differences, atypical speech, or dyslexic spelling patterns (Guo et al., 2019). However, as with native language/language proficiency, there has been little work on algorithmic bias in education connected to student disability status. Paquette and colleagues (2020) note that "very few" papers published in educational data mining 2015-2019 consider disability status; many of the papers in the field that do consider disabilities are focused on detecting whether a student has a disability (e.g. Käser Jacober, 2014; Klingler et al., 2017) rather than whether another algorithm or model is biased. However, at least two papers do investigate algorithmic bias involving disability in education. The use of speech recognition in educational assessment was evaluated by Loukina and Buzick (2017), who found that the accuracy of the SpeechRater system (used in a test of English language fluency) was much lower for students who appeared to have a speech impairment (according to notes taken by test administrators) but did not request accommodations, than for students who did not appear to have an impairment or students who requested accommodations for a speech impairment. Also, Riazy et al. (2020) investigate whether course outcome prediction is impacted by whether a student has a disability (any self-reported disability). They found that students with disabilities performed more poorly in the course and that algorithms applied to the data generally over-predicted success for students with disabilities, but evidence for different quality of prediction was fairly weak.

Urbanicity

A small number of studies have investigated algorithmic bias involving student urbanicity. Ocumpaugh and colleagues (2014) studied the effectiveness of interaction-based automated detectors of student confusion, frustration, boredom, and engaged concentration, trained on urban, suburban, rural, and a combined population. They found that the detectors were generally more effective for new students within the population they were trained on than for other populations; a detector developed on a combined population was more effective for urban and suburban students than for rural students. The authors noted that this was not simply due to race or socioeconomic status, since the suburban and rural students had a similar racial profile, and the urban and rural students had a similar socioeconomic profile.

Contrastingly, however, Samei and colleagues (2015) tested detectors that recognized attributes of student classroom questions, looking at whether models trained on data from urban learners worked in non-urban settings, and vice-versa. They found no degradation in model performance.

Parental educational background

Kai et al. (2017) differentiated performance between students whose parents had attended college and students whose parents did not attend college, in a model predicting student retention in an online college program. They found that their model had very good performance for both groups, but that it was better for the more at-risk group, students whose parents did not attend college. Yu and colleagues (2020) studied prediction of undergraduate course grades and average GPA and found that students who were first-generation college students were inaccurately predicted to perform worse than other students if educational background was included in the model, but that if only clickstream and survey data were included, the models became more fair for these students. Yu and colleagues (2021) studied prediction of college dropout, finding that

their models have worse accuracy and true negative rates for first-generation residential students, but better recall, regardless of whether protected attributes are included in the models

Socioeconomic status

Yudelson et al. (2014) trained and tested models across schools with high, medium, and low proportions of students eligible for free or reduced-price lunches, a common proxy for socioeconomic status in the United States. They compared between these groups in predicting a complex dependent measure that integrated several dimensions of student performance into a single variable. They found that models trained on schools with high proportions of low-SES students did not function as well in other schools, but that models trained in other schools functioned equally well for schools with high proportions of low-SES students. The differences seen, however, were small in magnitude. Yu and colleagues (2020) studied prediction of undergraduate course grades and average GPA and found that students who were from less wealthy backgrounds were inaccurately predicted to perform worse than other students if personal background was included in the model, but that if only clickstream and survey data were included, the models became fairer for these students. Yu and colleagues (2021) studied prediction of college dropout, finding that their models have worse accuracy and true negative rates for residential students with high financial needs, but better recall, regardless of whether protected attributes are included in the models

International students

Yu and colleagues (2020) studied prediction of undergraduate course grades and average GPA and found that international students were inaccurately predicted to perform worse than other students if personal background was included in the model, but that if only clickstream and survey data were included, the models became fairer for these students.

Military-connected status

Baker, Berning, and Gowda (2020) studied models predicting high school graduation and SAT score, training on students who were not military-connected (a student with a parent or close family member in the military) and testing on students who were military-connected, and vice-versa. They found evidence for moderate degradation in model quality when models were used in this fashion.

Table 2 Studies of algorithmic bias involving less frequently studied populations and subpopulations

	Native Language / Dialect	Disability	Urbanicity	Parental Education	Socio-economic Status	International Students	Military-connected Status
Ocumpaugh et al., 2014			X				
Yudelson et al., 2014					X		
Samei et al., 2015			X				
Kai et al., 2017				X			
Loukina & Buzick, 2017		X					
Naismith et al., 2018	X						
Loukina et al., 2019	X						
Baker at al., 2020							X
Riazy et al., 2020		X					
Yu et al., 2020				X	X	X	

3.3 Summary and discussion

Across categories, the findings of these studies seem to suggest that models trained on one group of learners perform more poorly when applied to new groups of learners. This is not universally true -- for example, there have been conflicting results for urban/rural learners, and the studies conducted across several nationalities often find different nationalities being disadvantaged in different analyses. But in aggregate, the findings suggest that it is problematic to ignore group differences when applying models. The simple expedient of collecting a diverse sample, and training on all students, seems to provide benefits in some cases. It may be that emerging methods for fairness-aware machine learning will lead to considerable improvements, once a representative sample is collected. Even if these methods are highly successful, we will also need to figure out how many members of an underrepresented group are necessary for a combined model to be valid, which remains a challenge in machine learning (see discussion and example in Slater & Baker, 2018). The trend in machine learning over the last few decades has largely been to consider ever-larger data sets rather than minimum data set sizes needed (Jiang et al., 2020). While not discounting the "unreasonable effectiveness of big data" (Halevy et al., 2009), we note that it is still necessary to determine how many learners of a specific group need to be in a training set (or a separate model's training set) before the model can generally be expected to be reliable for that group.

Another factor that is quickly apparent in looking across studies is the idiosyncracy of the categories that have been studied. Three core categories have received most of the attention from researchers: race/ethnicity (but not indigenous learners), gender (but not non-binary or

transgendered students), and nationality (for a small number of nationalities; in terms of learners' current locations). A handful of other categories have been discussed in one or two papers. The list of categories that have been studied seems idiosyncratic. To some degree it is based on convenience -- U.S. census categories are relatively likely to be collected, and a learner's current national location is likely to be known. To some degree it is based on the categorizations that are societally or politically important. To some degree it is based on the biases in what students even make it into the samples -- this may explain to a large degree why indigenous learners are omitted.

Even within the categories that have been relatively heavily studied, there is still considerable idiosyncrasy in the contexts where these categories are studied. For instance, despite the existence of large multi-national datasets involving MOOCs and considerable recent research using MOOCs as a context for conducting research on the differences between learners in different countries (Reich, 2015; Kizilcec & Brooks, 2017), MOOCs have not yet become a widely-used context for studying algorithmic biases involving national difference. Many of the findings discussed above were inconsistent across different studies. It is not yet clear whether this is simply due to noise and random factors, or whether some differences matter more in specific contexts than in other contexts. Fully understanding not only *which* categories matter, but what their characteristic manifestations are in different contexts, will need to wait until a much larger number of studies have occurred, conducted across a range of contexts.

It is not immediately obvious why some categories have been studied and other categories have not been studied. However, there seem to be effects showing up for a range of groups, suggesting that algorithmic bias likely impacts other groups as well. A broader range of groups need to be more explicitly studied. For instance, children of migrant workers experience many of the same challenges that military-connected students do, such as high personal mobility and concerns about the safety of family members abroad, but have not been studied. Religious minorities have not been studied. Age has not been studied as a factor in undergraduate courses, graduate courses, or professional learning.

Even when groups have been studied, they are often considered in an overly simplified fashion. Large and highly internally diverse groups are currently treated as a single entity. Why are Latinx learners or Spanish-speaking learners treated as monolithic groups, given the large amounts of variance in each of these categories? Why are all learners from China -- a highly-diverse country -- treated as being the same? Why are all Asian-Americans treated the same, from Japanese-Americans to Cambodian-Americans to Indian-Americans? The categories we typically work with come from political distinctions (Strmic-Pawl et al., 2018), not cultural or scientific distinctions. While detailed sub-division of groups may be impractical in small data sets, educational data sets used in AIED increasingly scale to tens or hundreds of thousands of learners. As such, many of the larger data sets studied in the field would be large enough to analyze differences between groups currently labeled together, if more fine-grained labels had been collected. Some distinctions may be infeasible to make in smaller data sets (true even for widely-used census categories), but documenting the differences between groups labeled together, where possible, can determine which labels are generally too coarse-grained, and what contexts and applications it is important to sub-divide them for.

The astute reader -- or at least a reader who has not grown up in the United States -- will also note the intense American focus of research discussed here. What about Cockneys, Northerners

(UK), Travellers, or for that matter the difference between Irish, Scottish, Welsh and English learners? What about learners living in different European countries? What about urban and rural learners in Brazil or China? What about Pakistani descendants living in the UK, Kurdish descendants living in Scandinavia, ethnic minorities in Russia, and members of the Portuguese diaspora? What about linguistic minorities in Spain, France, and Italy? The American focus of the research presented here is a reflection of where this research is currently occurring. Almost every single paper we were able to find that studies algorithmic bias in a specific group of learners, based on an externally identifiable non-malleable characteristic, for a specific algorithm, involved authors working in the United States. Fascinatingly, relatively few of these authors were born in the United States. It is imperative that this research extend beyond the United States.

4. From unknown bias to known bias; from fairness to equity

In the previous sections, we outlined the published literature on evidence for specific algorithmic biases in education. As the summary of that literature indicates, there is not just evidence for algorithmic bias -- there is evidence for several algorithmic biases. But this evidence is highly sparse -- in several cases, there is only a handful -- or even just one -- cases where a specific algorithmic bias has been documented.

We can posit a progression in a field's efforts to address bias. First, must come *unknown bias*-- we do not know that a problem exists, or perhaps we do not know its extent or its exact manifestations. When the field does the research to better understand the problem, over time it moves to a stage of *known bias* -- we have a working understanding (likely imperfect) of how serious a problem is and what situations it emerges in. It is difficult to address a problem until it is reasonably well-understood, otherwise our efforts may be directed to the wrong manifestations of the problem, or to its symptoms rather than its causes. Even if we cannot discover all biases, or fully understand a bias we are aware of, we can do much better at discovering and understanding bias than the current state of the field (shown in section 3).

But, eventually, by understanding a specific bias and engaging in efforts to fix it (perhaps theoretically-driven, perhaps more trial-and-error), we can get to a point where we come closer to *fairness*. Even if it is mathematically impossible to simultaneously optimize for all definitions of fairness (see discussion in section 2.3), we again can improve considerably on the current state of the field, where most algorithms do not even consider fairness, even informally.

Fairer algorithms can be a step towards designing for *equity* -- creating a better world, with equal opportunity for all (see discussion in Holstein & Doroudi, in press). Equity for all does not imply ignoring the inequities that specific groups have suffered, often multi-generationally. Working towards equity necessarily implies focusing on the biggest problems -- the biggest failures of fairness and the biggest inequities -- first. But it also depends on understanding all the places that inequity hides in the shadows, the inequities that we may fail to see because of the biases baked into our assumptions and societal narratives. It is imperative on all of us to look for evidence of unknown or poorly-understood biases, and not simply to assume that the problems widely known today are the only problems worth tackling. If we analyze the attention given to algorithmic bias in the last few years, particularly outside of education, we see an intense focus on fairness -- developing metrics and algorithms for fairness. But it will be hard to achieve fairness if we cannot move from unknown biases to known biases. Perhaps the strongest message of this paper is how little we as a field know today about algorithmic bias in education, and how much more we have to learn.

4.1 What obstacles stand in the way of research into unknown algorithmic biases?

Given the importance of revealing and understanding unknown biases, the next question is -- what can we do to make more progress and faster progress in this area?

Unknown biases come in multiple types. One type is a bias that is completely unknown -- we do not know that algorithmic biases exist for a specific group. There are likely to be several such biases out there waiting for our field to discover them -- given the aforementioned sparsity of the data on algorithmic bias in education, it seems highly improbable that we have discovered all important biases. This first type of unknown bias is very concerning. Without understanding which groups are at risk of algorithmic bias, even very well-intentioned attempts to resolve algorithmic bias may miss at-risk groups of students.

For example, systems like The Generalizer (Tipton, 2014) are often used to select samples for large-scale studies. The Generalizer has the goal of helping researchers select populations that are representative, by examining the degree to which the sample selected is representative of the target population in terms of pre-selected variables. However, those variables encode a perspective on which demographic and group variables matter; not all of the variables that appear to be associated with algorithmic bias in education (according to our review) are available in The Generalizer.

However, a second type of unknown bias is equally serious -- cases where we know that bias exists, in general, but not how it manifests itself or where it occurs. We would argue, based on the evidence provided in section 3, that this is true even for very well-known problems such as racism and sexism. The scattered and sometimes inconsistent evidence on algorithmic bias, even for these groups, suggests that our field still has a great deal to learn.

Perhaps the first and seemingly most obvious step is to collect more data on group membership, when data is already being collected. As Paquette and colleagues (2020) note, most research in educational data mining publication venues does not even mention learner demographics at an aggregate level. If we do not collect data on group membership, we cannot analyze our data sets for algorithmic bias.

However, even this simple step imposes challenges -- both concrete and abstract. First of all, reasonable concerns about privacy raise the possibility of risk stemming from collecting this type of data. For instance, it may be possible to re-identify a student from their demographic data (cf. Benitez & Malin, 2010); a classroom may only have one female indigenous student, and therefore reporting this information creates a serious privacy risk. There are methods that can be used to reduce this risk, discussed in the next section, but the risk is hard to entirely eliminate in a small data set.

Second, some forms of group data may be restricted in some contexts, due to legal or regulatory requirements. For example, many institutional review boards (or similar regulatory organizations in other countries) consider demographic data to be higher-risk and therefore create greater hurdles for collecting this data, creating an incentive for researchers to ignore issues of algorithmic bias. So too, differing protections on data on student disability status between countries have an impact on whether researchers are able to investigate algorithmic bias (or even design) related to student cognitive disabilities.

Several other challenges prevent the deeper research into unknown biases for groups known to be at risk of algorithmic bias in general. One is economic. There is a strong commercial disincentive for the developers of learning systems to reveal the algorithmic biases in their systems. Just collecting the data necessary to investigate algorithmic biases presents risks. Even in cases where the privacy risk for any specific student seems logically very small, there are many who will enthusiastically critique an apparent data risk, particularly if it involves a company or individual disliked for other reasons. Going further and publishing evidence that one's system has algorithmic biases -- even in an effort to fix those biases -- risks public critique by journalists, community members, and academics more eager to decry bias than fix it. It can potentially even risk lawsuits. As such, there is a strong "keeping one's head down" incentive not to collect or analyze group data.

Simultaneously, educational effectiveness is often treated as universal -- for example, clearinghouses such as the What Works Clearinghouse and Evidence for ESSA treat curricula as having evidence for efficacy or not having evidence for efficacy, rather than effective or ineffective for specific contexts or groups of learners. As such, in many cases, publicly investigating algorithmic bias currently may present more risk for commercial learning system developers than benefit. Fortunately, recently some school districts have begun to ask for evidence not only that a system works in general but also that it works for students similar to their current students (Rauf, 2020). This may create a countervailing incentive for developers to collect data and investigate algorithmic bias in their systems, or to partner with external researchers in doing so.

Another important possibility is that some algorithmic bias in education may be intersectional (Crenshaw, 1991) -- i.e. creating a specific impact based on multiple group memberships. For example, Black female students may be impacted differently than Black male or non-Black female students. There has been insufficient research into intersectionality in educational work on algorithmic bias, and indeed in algorithmic bias more generally (Cabrera et al., 2019). While there may not always be enough data to study intersectionality, where it is possible, it is an important next step.

Ultimately, there is a lot of work to be done to understand algorithmic biases in education that are currently unknown or incompletely understood. Algorithmic, metrics-based work to expose bias at downstream locations of the ML pipeline will have limited impact on the equity of educational opportunity and outcomes for students, without having the right data in place. This will require extensive effort to re-engineer incentives, eliminate barriers, and mitigate challenges that prevent this work. Only by addressing these challenges can we achieve the necessary steps of expanding the distribution of data collection and research efforts to populations being impacted by algorithmic bias. Only by addressing these challenges can we move towards fully understanding how population-specific factors (both well-known demographic factors and less-understood factors) impact the effectiveness of educational algorithmics and the interventions and system responses that build upon them.

In the next section, we take up the issue of how these obstacles -- and other obstacles to resolving algorithmic bias -- can be more effectively addressed.

4.2 How should researchers and educators address these obstacles?

There are a number of steps that can be taken to address these obstacles. While no one of these steps can by itself eliminate algorithmic bias -- it is not clear that algorithmic bias *can* be entirely eliminated -- any one of these steps has the potential to drive progress in reducing the degree to which bias is present in the field's algorithms (for a summary of recommendations see Table 3).

Improving data collection

Our first area of recommendation -- likely unsurprising to anyone who has read through this paper until this point -- is to improve data collection. As our review above notes, the majority of the work on algorithmic bias, including in education, has focused on the stages of the process where an algorithm is developed and/or evaluated. However, it is difficult for any algorithmic or metric-based approach to make a positive impact if appropriate data is not available to the algorithm.

Towards getting the right data in place, we recommend that researchers collect extensive demographic data on learners whenever possible. It is impossible to evaluate if a model is fair towards a specific group of learners if we do not know which learners are in that group. These efforts should work to collect traditional census categories, absent in many data sets used in our field (Paquette et al., 2020), but this is a minimum -- as our review has shown, many other categories are associated with algorithmic bias. It is too early to list out all categories that should be included in an effort of this nature -- there are too many unknown unknowns, and more research is needed. However, it is not too early to focus more energy on our known unknowns and collect data about student membership in identity categories that already appear to be associated with algorithmic bias.

Based on the evidence thus far, we strongly recommend that whenever possible researchers collect data on gender, race, ethnicity, and national origin. There is not yet sufficient evidence about disability status, dialect, socioeconomic status, urbanicity, native language (and second-language learner status), national region, parental education background, military-connectedness, or migrant work to make broad recommendations, but that is due to the lack of research on how algorithmic bias manifests for these aspects of identity. We suspect that -- given research -- these categories will emerge as important to collect data on as well, and therefore recommend that researchers collect data on these categories as well, whenever possible.

Of course, as discussed in section 4.1, our recommendation to collect richer identity data does have the drawback of increasing risk around privacy and increased regulatory challenges. Approaches such as data obfuscation (Bakken et al., 2004), providing researchers the ability to use but not view variables (Gardner et al., 2018), legal agreements around data re-identification (ASSISTments Project, 2014), can mitigate these risks to a degree. Encouraging regulators and institutional review boards (or other privacy officers) to balance the risks of privacy violations with the risks of algorithmic bias will also be highly important. Once an algorithm development project has collected data on learners' identities, a next step will be to identify gaps in the data set's representativeness and address those gaps through additional data collection. As such, the field needs to do additional work to create practices for making sure training sets are representative and to address underrepresentation of key groups in data sets. A key step in this process will be to determine how much data is actually needed for subpopulations within the data set, in order to have reasonably high confidence that the model will work for new data in these subpopulations. There are not yet methods for selecting sample sizes for most types of machine

learning that have the reliability of statistical power analysis (Kraemer & Blasey, 2015), but methods such as adversarial filtering -- which down-sample overrepresented groups rather than collecting new data for underrepresented groups (Le Bras et al., 2020) -- may provide insights that can eventually lead to guidelines on minimum data set sizes for underrepresented groups. Of course, in increasing efforts to collect data from underrepresented learners, it will be important not to impose undue burdens on learners (or teachers and administrators) in underrepresented contexts (Chicago Beyond, 2019).

Another important area where data collection can be improved is in avoiding (or reducing) bias during the labeling process. Both researcher-generated labels (such as classroom observations of student engagement) and labels coming from external sources (such as school data on violence or grades) can themselves be biased. As discussed in section 2.4, any label based on a human's judgment -- even self-report -- is at risk of being biased in this fashion. Therefore, it may be beneficial to replace subjective judgments with other training labels, where possible. However, the solution is not to abandon subjective judgment training labels entirely. Indeed, many constructs -- such as 21st-century skills and affect -- can currently only be detected based on subjective judgment training labels. Okur et al.'s work (2018) suggests that members of the group being labeled will provide more accurate and less biased training labels than members of other group. Where possible, data labeling should therefore be conducted by members of the group being studied. If this is infeasible, evidence for biases in the training data can be addressed by creating a distilled variable that partials out the effects of group variables.

Improve tools and resources

The last several years have seen a considerable amount of interest in different algorithms and metrics for addressing algorithmic bias, both in education and more broadly. However, the ideas represented in these papers have not emerged into widespread use within educational research and development around algorithms. Part of the challenge is that applying many of the algorithms and metrics that are available requires a developer to either implement these algorithms themselves, or to use often poorly-documented and buggy software packages.

Better software tools would therefore be a useful step towards increasing uptake of best practices. Hajian and colleagues (2016) make a distinction between tools for algorithms that increase fairness/reduce bias and tools that evaluate if bias is present. There is an increasingly large number of algorithms for increasing fairness, and research is proceeding quickly in this area. It is not yet clear which algorithms will be most effective and useful in education. It might therefore be premature to create standard tools for algorithms, especially considering the high amounts of effort needed to create scalable, production-quality algorithm packages.

However, the situation for metrics for evaluating if an algorithm is biased is somewhat different. Metrics are typically considerably less complex than algorithms and are therefore more straightforward to implement. And having several metrics available facilitates evaluation and decision-making around algorithms. As Mitchell and colleagues (2021) note, considering a range of metrics makes it possible to consider the trade-offs around different algorithms and ways of applying them, as well as making assumptions explicit. Therefore, it is both feasible and desirable -- right now -- to create packages in Python and R that can be used by researchers to obtain a standard set of metrics around algorithmic bias and fairness. Efforts are already underway to create domain-general packages for algorithmic bias/fairness metrics, and their developers welcome external contributions (Bellamy et al., 2019). Thus, we call for an effort to

build on these domain-general packages to create educational algorithm focused packages. This effort would focus on providing functionality for researchers and developers to quickly obtain a standard set of metrics.

A second potentially useful resource would be the creation of "reference data sets" for use by researchers interested in algorithmic aspects of fairness, within education. Reference data sets have been an essential part of progress in a range of domains, from computer vision (Soomro et al., 2012), to computational linguistics (Bird et al., 2008), and -- in educational research -- for refining models that infer student knowledge (Selent et al., 2016; Stamper & Pardos, 2016). Researchers of algorithmic bias in other domains have recommended the creation and curation of high-quality reference data sets (Paullada et al., 2020). A reference data set can be used to compare algorithms to reduce algorithmic bias, to compare metrics for evaluating algorithmic bias, and as a model for future data set developers to compare their own efforts to. Within education, reference data sets should contain a wide variety of demographic variables (as detailed in the previous sub-section), have a variety of label variables across data sets (both macro-outcomes such as school dropout or course failure, and finer-grained variables such as student affect), and should have clear evidence for algorithmic bias.

Openness and Incentive Structures

A key step towards fixing algorithmic bias will be to shed greater light on where it is occurring -- reporting it and making the community aware of where the problems are. Given the incentives *against* openly reporting algorithmic bias, discussed above, we need steps to re-align incentives in favor of reporting algorithmic bias.

One such step would be for scientific journal editors and conference committees to adopt guidelines for analyzing algorithmic bias in education. Under these guidelines, all papers that report the use of prediction algorithms on authentic data sets collected after a selected year (announced in advance) would be required to conduct analyses to investigate whether algorithmic bias is present. Getting the details of this requirement right (which analyses? which metrics? which groups? how to apply this standard evenly worldwide? how to avoid a rush of papers to the journals which do not adopt standards?) would require concerted effort among many stakeholders, but would exert a powerful influence on practice in academia and the commercial organizations that value the reputational and other benefits of academic publishing.

This effort could be coordinated with ongoing efforts at school districts and local education agencies to create standards for demonstrating effectiveness for the local student population (Rauf, 2020), and could be combined with an effort to convince organizations such as the What Works Clearinghouse and Evidence for ESSA to consider evidence for generalizability instead of treating educational effectiveness as universal. In these efforts, more targeted to the general public, it might be worthwhile for requirements to include straightforward evidence on whether there is algorithmic bias, as well as more specialized technical demonstrations.

Beyond this, these same organizations could move towards creating guidelines and expectations for opening access to data sets and algorithms for inspection and critique, as is seen in the biomedical field and in journals like *Science* and *Nature*. However, companies have legitimate reasons to want to protect their intellectual property, developed at considerable cost and effort. One compromise might be for education agencies to require companies to allow access to the algorithms to specific designated research scientists who would conduct algorithmic bias reviews,

the practice used by governmental agencies in demonstrating the effectiveness of new medicines (e.g. Ciociola et al., 2014).

It may be possible to drive greater openness and attention to these problems via positive incentives as well as the more negative pressure of guidelines. For instance, journal editors and conference organizers in education could create special issues and workshops that provide a forum and an opportunity to publish emerging research around algorithmic bias. These forums would be particularly beneficial if they created an opportunity for preliminary work around unknown unknowns -- creating a space for first publication of evidence for algorithmic bias impacting groups that have not been previously studied.

Another form of positive incentive is funding. Funders, both governmental and non-governmental, could create funding streams (or possibilities within existing funding streams) for research on specific, concrete forms of algorithmic bias in education. It also may be possible for non-profit and advocacy organizations to create awards that recognize accomplishments in revealing and addressing algorithmic biases, such as a public award for the commercial organization that has done the most this year to address inequities stemming from algorithmic bias.

Broaden the Community

One of the key ways to reduce algorithmic bias in education is to broaden the community of people who are working on solving these problems. Educational researchers and data scientists do not need to solve these problems on their own and will be less successful if they try to. Algorithmic bias in education ultimately impacts society in general, and involving teachers, school leaders, parents, students, employers, policymakers, and community organizers in the process of thinking about educational algorithms has several potential benefits.

There have been calls to broaden the discussion around the design and use of algorithms (Cramer et al., 2019; Mitchell et al., 2021), including in education (Holstein & Doroudi, in press), but these calls have not been followed up on to the degree necessary, either in educational domains or in the consideration of algorithmic bias more broadly. As Olteanu and colleagues (2019) state, “The interpretation and assessment of results are too often done by data experts, not by domain experts.” Members of communities being affected can always do a better job of advocating for their perspective than well-meaning outsiders, and -- indeed -- well-meaning outsiders often fail to fully understand the issues at hand or the constraints on a successful solution (L. T. Smith, 2013). As such, involving community members in the entire process of algorithmic development and use can support more accurate interpretations of data, produce better hypotheses for the causes of the phenomena being observed, and lead to designs and interventions more likely to be found acceptable by the teachers and school leaders implementing them and the students and families who are affected. Fuller inclusion of the people who will use and be impacted by algorithms may also lead to better designs for the implementation and use of algorithms, avoiding cases where unfairness and bias stem not from an algorithm itself, but from how it is used.

These efforts are unlikely to be fully successful unless algorithms and their properties are effectively communicated to these community members. Community members -- and non-data scientists in general -- cannot provide useful recommendation and oversight if they are not given the appropriate information on algorithms and their properties. It is not just a matter of making

the information available, it must also be understandable (Howley, 2018). Hence, progress in explainable artificial intelligence and interpretable artificial intelligence (Doran et al., 2018) and in research on explanation methods that make complex algorithms understandable (Zhou et al., 2020) will be essential to addressing the challenges surrounding algorithmic bias. Modern explainable AI methods make it possible to see the impact of a specific variable on prediction (Lundberg et al., 2020), making it easier to see not just where the biases are, but what variables might be causing the biases, towards developing better, less-biased algorithms. These efforts will also depend on bringing in a wider diversity of voices through the entire process of developing an algorithm and using it in practice. Inclusion cannot simply be asking for opinions at one stage in the process. Doing so may be perceived as an insincere attempt to obtain buy-in rather than an attempt to improve the algorithm, and those judgments may not be wrong. As noted by M. K. Lee and colleagues (2019), not only the outcome, but the process that derives the outcome determines perceptions of bias. A full and open inclusion of community members may suggest better variables to use, better consideration of trade-offs, and even may be able to mitigate the negative impacts of a flawed algorithm through choosing better ways to use its outputs and better interventions (see discussion in Mitchell et al., 2021). These partners may also be able to identify additional sources of bias in a learning technology beyond just the algorithm. As noted in several articles (i.e. Arroyo et al., 2013; Baker et al., 2019; Finkelstein et al., 2013; Mayfield et al., 2019; Melis et al., 2009; Woolf et al., 2010), considerable bias in education can come from how a learning interaction is designed.

Although the primary sense in which we make this recommendation is in broadening the community that considers this issue by adding members of impacted communities, another sense in which the community should be broadened is in terms of intellectual perspectives. As feminist theory (D'ignazio & Klein, 2020), sociocultural theory (Ferrero & Barujel, 2019), and critical race theory (Hanna et al., 2020) begin to influence and improve data science, we need the ideas from these communities in AIED research as well. Ideas from these communities have influenced this article and bringing these voices into our community to a greater degree will enhance our field's ability to go from bias to fairness, and from fairness to equity.

Finally, in terms of broadening voices, it is essentially to bring a broader collection of voices not just into academic discussion of algorithmic bias, but into industrial work around educational algorithms as well. Corporate voices have been prominent in discussions of algorithmic bias in broader computer science discourse (Blodgett et al., 2020; Cramer et al., 2019; Gebru et al., 2018; Mitchell et al., 2021), but have been relatively less prominent in education. Building capacity in the developers of learning systems that use educational algorithms, including the inclusion of diverse voices and ideas in these organizations, will be essential to reducing algorithmic bias across the educational landscape.

Table 3 General strategies and specific recommendations for reducing algorithmic bias in education

	Recommendations
Improve Data Collection	• Prioritize collection of data on gender, race, ethnicity, and national origin. • When possible collect data on disability status, dialect, socioeconomic status, urbanicity, native language (and second-language learner status), national region, parental education background, military-connectedness, and migrant status. • Create infrastructure to reduce privacy risks while still using fuller information. • Encourage regulators and IRBs (or other privacy officers) to balance the risks of privacy violations with the risks of algorithmic bias. • Create practices for making sure training sets are representative. • Avoid (or reduce) bias during data labeling.
Improve Tools and Resources	• Create standard, education-specific, packages to calculate bias metrics and to conduct bias audits. • Create reference datasets for testing new approaches.
Create Structures to Incentive Openness	• Adopt journal and conference guidelines requiring analysis for algorithmic bias. • Incorporate evidence of generalizability when demonstrating effectiveness. • Consider options for opening data sets to inspection and critique. • Create or facilitate journal special issues and publication opportunities. • Encourage funding for research that investigates algorithmic bias in education.
Broaden the Community	• Involve members of communities potentially impacted by algorithms throughout the entire process of algorithm development and use. • Expand efforts to make artificial intelligence more explainable and interpretable.

5. Conclusion

In this article, we have reviewed the contemporary problem of algorithmic bias in education. We contextualize this within the broad scope of literature on algorithmic bias, across sectors, but focus on the specific manifestations of this problem in education. We discuss key theoretical perspectives on algorithmic bias and then turn to reviewing the still very sparse literature on how specific groups of learners are being impacted by algorithmic bias. This review reveals that there are a great number of "known unknowns" in the study of algorithmic bias in education -- areas where we know that our current knowledge is insufficient -- but also exposes the possibility that there may also be many "unknown unknowns" -- groups that we do not even realize are being severely impacted. Foremost among these "unknown unknowns" is the lack of empirical work on how algorithmic bias impact specific groups of learners outside the United States, at a finer grain-size than the national level.

This calls for an ambitious program of research, across our field, to study the known unknowns and reveal the unknown unknowns. We discuss the challenges currently present to the successful completion of this program of research, and offer potential solutions to these challenges, focusing on four potential areas of solution: improved data collection, improved tools and

resources, increasing openness and improving incentive structures, and broadening the community of stakeholders who are included in our field's efforts to address algorithmic bias.

We do not pretend that these steps will at last resolve the deep, fundamental inequities in educational opportunity and practice around the world. We hope for a world with fewer unknown biases, where known biases are better documented, where known biases have been reduced in magnitude and impact, and where the world is more equitable. Getting there will be difficult. However, by working together to reduce and find ways to mitigate algorithmic bias, we can at minimum reduce the risk of worsening the situation further -- and we may be able to make a concrete positive difference.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359301

Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. *Proceedings of the 12th International Conference on Educational Data Mining*, 488–491.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing%0A

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender Differences in the Use and Benefit of Advanced Learning Technologies for Mathematics. *Journal of Educational Psychology*, *105*(4), 957–969. https://doi.org/https://doi.org/10.1037/a0032748

ASSISTments Project. (2014). *ASSISTmentsData: Terms of Use for Using Data*. Retrieved January 7, 2021, from https://sites.google.com/site/assistmentsdata/termsofuseforusingdata

Baker, R. S. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining*, *11*(1), 1–17. https://doi.org/10.5281/zenodo.3554745

Baker, R. S., Berning, A., & Gowda, S. M. (2020). Differentiating Military-Connected and Non-Military-Connected Students: Predictors of Graduation and SAT Score. *EdArXiv*. https://doi.org/10.35542/osf.io/cetxj

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D. J., & Beck, J. E. (2006). Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392–401. https://doi.org/10.1007/11774303_39

Baker, Ryan S., Walker, E., Ogan, A., & Madaio, M. (2019). Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context*, *1*(1), 1–13.

Bakken, D. E., Rarameswaran, R., Blough, D. M., Franz, A. A., & Palmer, T. J. (2004). Data obfuscation: anonymity and desensitization of usable data sets. *IEEE Security & Privacy*, *2*(6), 34–41. https://doi.org/10.1109/MSP.2004.97

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, *17*(2), 169–177. https://doi.org/10.1136/jamia.2009.000026

Benner, K., Thrush, G., & Isaac, M. (2019, March 28). Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says. *New York Times*. https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. https://doi.org/https://doi.org/10.1016/j.specom.2007.02.006

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., Lee, D., Powley, B., Radev, D. R., & Tan, Y. F. (2008). No Title. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 1755–1759.

Bireda, M. R. (2002). *Eliminating Racial Profiling in School Discipline: Cultures in Conflict*. Scarecrow Press.

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Blodgett, S. L., & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *ArXiv E-Prints*, arXiv:1707.00061. https://arxiv.org/abs/1707.00061

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April 13-17). *Considering fairness and validity in evaluating automated scoring* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA, United States.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, *25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *ArXiv E-Prints*, arXiv:2010.04053. https://arxiv.org/abs/2010.04053

Chicago Beyond. (2019). *Why am I always being researched? A guidebook for community organizations, researchers, and funders to help us get from insufficient understanding to more authentic truth.* Chicago Beyond. https://chicagobeyond.org/researchequity/

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, 726–731.

Ciociola, A. A., Cohen, L. B., Kulkarni, P., & FDA-Related Matters Committee of the American College of Gastroenterology. (2014). How drugs are developed and approved by the FDA: current process and future directions. *The American Journal of Gastroenterology*, *109*(5), 620–623. https://doi.org/10.1038/ajg.2013.407

Cole, N. S., & Zieky, M. J. (2001). The New Faces of Fairness. *Journal of Educational Measurement*, *38*(4), 369–382. https://doi.org/10.1111/j.1745-3984.2001.tb01132.x

Cramer, H., Holstein, K., Vaughan, J. W., Daumé, H., Dudik, M., Wallach, H., Reddy, S., & Jean, G.-G. [The Conference on Fairness, Accountability, and Transparency (FAT*)]. (2019, February 23). *FAT* 2019 Translation Tutorial: Challenges of incorporating algorithmic fairness* [Video]. YouTube. https://youtu.be/UicKZv93SOY

Crawford, K. [The Artificial Intelligence Channel]. (2017, December 11). *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford* [Video]. YouTube. https://youtu.be/fMym_BKWQzk

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1300.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, *8*(2), 71–82. https://doi.org/10.1111/j.1745-3984.1971.tb00908.x

Dieterle, E., Dede, C., & Walker, M. (under review). The cyclical ethical effects of using artificial intelligence in education. Manuscript under review.

D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Doran, D., Schulz, S., & Besold, T. R. (2018). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *CEUR Workshop Proceedings*, *2071*. https://openaccess.city.ac.uk/id/eprint/18660/

Dorans, N. J. (2010). Misrepresentations in Unfair Treatment by Santelices and Wilson. *Harvard Educational Review*, *80*(3), 404–413.

Doroudi, S., & Brunskill, E. (2019). Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335–339. https://doi.org/10.1145/3303772.3303838

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Ferrero, F., & Gewerc Barujel, A. (2019). Algorithmic Driven Decision-Making Systems in Education: Analyzing Bias from the Sociocultural Perspective. *2019 XIV Latin American Conference on Learning Technologies (LACLO)*, 166–173. https://doi.org/10.1109/LACLO49268.2019.00038

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013). The effects of culturally congruent educational technologies on student achievement. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 493–502). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_50

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, *33*(4), 111–117. https://www.muse.jhu.edu/article/645268

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018). MORF: A Framework for Predictive Modeling and Replication At Scale With Privacy-Restricted MOOC Data. *2018 IEEE International Conference on Big Data (Big Data)*, 3235–3244. https://doi.org/10.1109/BigData.2018.8621874

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234. https://doi.org/10.1145/3303772.3303791

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé Hal, I. I. I., & Crawford, K. (2018). Datasheets for Datasets. *ArXiv E-Prints*, arXiv:1803.09010. https://arxiv.org/abs/1803.09010

Green, B. (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. https://doi.org/10.1145/3351095.3372869

Green, B., & Hu, L. (2018, July 10-15). *The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning* [Conference presentation]. The Debates Workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden.

Green, B., & Viljoen, S. (2020). Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31. https://doi.org/10.1145/3351095.3372840

Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., & Morris, M. R. (2019). Toward fairness in AI for people with disabilities: A research roadmap. *arXiv preprint arXiv:1907.02227*. https://arxiv.org/abs/1907.02227

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. https://doi.org/10.1145/2939672.2945386

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 501-512).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning -- What is it Good for? In A. Saffiotti, L. Serafini, & P. Lukowicz (Eds.), *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020)* (pp. 3–10). RWTH Aachen University.

Holstein, K. & Doroudi, S. (in press). Equity and artificial intelligence in education: Will "AIEd" Amplify or Alleviate Inequities in Education? Invited chapter in Porayska-Pomsta, K. & Holmes, W. (Eds.), *Ethics in AIED: Who Cares? Data, algorithms, equity and biases in educational contexts.* Milton Park, UK: Routledge Press.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3290605.3300830

Howley, I. (2018) If an algorithm is openly accessible, and no one can understand it, is it actually open? In *Artificial Intelligence in Education Workshop on Ethics in AIED 2018.*

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 431–437.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, *83*(6), 1053–1071. https://doi.org/10.1037/0033-2909.83.6.1053

Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)Fairness: Lessons for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

James, R., Bexley, E., Anderson, M., Devlin, M., Garnett, R., Marginson, S., & Maxwell, L. (2008). *Participation and equity: a review of the participation in higher education of people from low socioeconomic backgrounds and Indigenous people*. http://hdl.voced.edu.au/10707/31488

Jiang, J., Wang, R., Wang, M., Gao, K., Nguyen, D. D., & Wei, G.-W. (2020). Jiang, J., Wang, R., Wang, M., Gao, K., Nguyen, D. D., & Wei, G. W. (2020). Boosting tree-assisted multitask deep learning for small scientific datasets. *Journal of Chemical Information and Modeling*, *60*(3), 1235–1244.

Johns, J., & Woolf, B. (2006). A Dynamic Mixture Model to Detect Student Motivation and Proficiency. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 163–168.

Kai, S., Andres, J. M. L. ., Paquette, L., Baker, R. S. ., Molnar, K., Watkins, H., & Moore, M. (2017). Predicting Student Retention from Behavior in an Online Orientation Course. *Proceedings of the 10th International Conference on Educational Data Mining*, 250–255.

Käser Jacober, T. (2014). *Modeling and Optimizing Computer-Assisted Mathematics Learning in Children* [Doctoral dissetation, ETH Zurich]. ETH Library. https://doi.org/10.3929/ethz-a-010265296

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3819–3828). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702520

Kizilcec, R. F., & Brooks, C. (2017). Diverse big data and randomized field experiments in MOOCs. *Handbook of Learning Analytics*, 211-222

Kizilcec, R. F., & Lee, H. (in press). Algorithmic Fairness in Education. Algorithmic Fairness in Education. In W. Holmes & K. Porayska-Pomsta (Eds.), *Ethics in Artificial Intelligence in Education*. Abingdon-on-Thames, UK: Taylor & Francis.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In C. H. Papadimitriou (Ed.), *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, pp. 43:1-43:23). Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Klingler, S., Wampfler, R., Käser, T., Solenthaler, B., & Gross, M. (2017). Efficient Feature Embeddings for Student Classification with Variational Auto-Encoders. *Proceedings of the 10th International Conference on Educational Data Mining*, 72–79.

Knight, W. (2019, November 19). The Apple Card Didn't "See" Gender—and That's the Problem. *Wired*.

Kraemer, H. C., & Blasey, C. (2015). *How Many Subjects?: Statistical Power Analysis in Research*. SAGE Publications. https://books.google.com/books?id=wMxuBgAAQBAJ

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*(1), 56–65. https://doi.org/10.1037/0021-9010.70.1.56

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., & Choi, Y. (2020). Adversarial filters of dataset biases. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 1078–1088.

Lee, H., & Kizilcec, R. F. (2020). Evaluation of Fairness Trade-offs in Predicting Student Success. *ArXiv E-Prints*, arXiv:2007.00088. https://arxiv.org/abs/2007.00088

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), Article 182. https://doi.org/10.1145/3359284

Li, X., Song, D., Han, M., Zhang, Y., & Kizilcec, R. F. (2021). On the limits of algorithmic prediction across the globe. *arXiv preprint arXiv:2103.15212*

Loukina, A., & Buzick, H. (2017). Use of Automated Scoring in Spoken Language Assessments for Test Takers With Speech Impairments. *ETS Research Report Series*, *2017*(1), 1–10. https://doi.org/https://doi.org/10.1002/ets2.12170

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–10.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). On the Applicability of ML Fairness Notions. *ArXiv E-Prints*, arXiv:2006.16745. https://arxiv.org/abs/2006.16745

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W. (2019). Equity Beyond Bias in Language Technologies for Education. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 444–460. https://doi.org/10.18653/v1/W19-4446

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv E-Prints*, arXiv:1908.09635. https://arxiv.org/abs/1908.09635

Melis, E., Goguadze, G., Libbrecht, P., & Ullrich, C. (2009). Culturally Adapted Mathematics Education with ActiveMath. *AI & Society*, *24*(3), 251–265. https://doi.org/10.1007/s00146-009-0215-4

Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, *9*, 70–89.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*. https://doi.org/10.1146/annurev-statistics-042720-125902

Naismith, B., Han, N.-R., Juffs, A., Hill, B., & Zheng, D. (2018). Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data. *Proceedings of 11th International Conference on Educational Data Mining*, 259–265.

O'Reilly-Shah, V. N., Gentry, K. R., Walters, A. M., Zivot, J., Anderson, C. T., & Tighe, P. J. (2020). Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *British Journal of Anaesthesia*, *125*(6), 843–846. https://doi.org/10.1016/j.bja.2020.07.040

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501. https://www.learntechlib.org/p/148344

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. *International Journal of Artificial Intelligence in Education*, *25*(2), 229–248. https://doi.org/10.1007/s40593-014-0034-8

Okur, E., Aslan, S., Alyuz, N., Arslan Esme, A., & Baker, R. S. (2018). Role of Socio-cultural Differences in Labeling Students' Affective States. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (pp. 367–380). Springer International Publishing. https://doi.org/10.1007/978-3-319-93843-1_27

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, *12*(3), 1–30. https://doi.org/10.5281/zenodo.4143612

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *ArXiv E-Prints*, arXiv:2012.05345. https://arxiv.org/abs/2012.05345

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, *13*(1), 3–29. https://doi.org/10.1111/j.1745-3984.1976.tb00178.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. https://doi.org/https://doi.org/10.1016/j.asw.2012.10.004

Ramineni, C., & Williamson, D. (2018). Understanding Mean Score Differences Between the e-rater® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE® General Test. *ETS Research Report Series*, *2018*(1), 1-31.

Rauf, D. S. (2020, October 2). *The New, Tough Expectations Education Companies Face on Race and Diversity*. Market Brief: Market Trends. https://marketbrief.edweek.org/market-trends/new-tough-expectations-education-companies-face-race-diversity/

Reich, J. (2015). Rebooting MOOC research. *Science*, *347*(6217), 34-35.

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)*, *1*, 15–25. https://doi.org/10.5220/0009324100150025

Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). How Mastery Learning Works at Scale. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 71–79. https://doi.org/10.1145/2876034.2876039

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., & Graesser, A. (2015). Modeling Classroom Discourse: Do Models That Predict Dialogic Instruction Properties Generalize across Populations? *Proceedings of the 8th International Conference on Educational Data Mining*, 444–447.

Santelices, M. V., & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review*, *80*(1), 106–134. https://doi.org/10.17763/haer.80.1.j94675w001329270

Selent, D., Patikorn, T., & Heffernan, N. (2016). ASSISTments Dataset from Multiple Randomized Controlled Experiments. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 181–184. https://doi.org/10.1145/2876034.2893409

Silva, S., & Kenney, M. (2018). Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon (1960-)*, *55*(1&2), 9–37. https://www.jstor.org/stable/10.2307/26545017

Slater, S., & Baker, R. S. (2018). Degree of error in Bayesian knowledge tracing estimates from differences in sample sizes. *Behaviormetrika*, *45*, 475–493. https://doi.org/10.1007/s41237-018-0072-x

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI & SOCIETY*, *35*(4), 1077–1078. https://doi.org/10.1007/s00146-020-01054-3

Smith, L. T. (2013). *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed Books. https://books.google.com/books?id=8R1jDgAAQBAJ

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *ArXiv E-Prints*, arXiv:1212.0402. https://arxiv.org/abs/1212.0402

Soundarajan, S., & Clausen, D. L. (2018). Equal Protection Under the Algorithm: A Legal-Inspired Framework for Identifying Discrimination in Machine Learning. *Proceedings of the 35th International Conference on Machine Learning*.

Stamper, J., & Pardos, Z. A. (2016). The 2010 KDD Cup Competition Dataset: Engaging the Machine Learning Community in Predictive Learning Analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16%0A

Strmic-Pawl, H. V, Jackson, B. A., & Garner, S. (2018). Race Counts: Racial and Ethnic Data on the U.S. Census and the Implications for Tracking Inequality. *Sociology of Race and Ethnicity*, *4*(1), 1–13. https://doi.org/10.1177/2332649217742869

Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. *ArXiv E-Prints*, arXiv:1901.10002. https://arxiv.org/abs/1901.10002

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, *56*(5), 44–54. https://doi.org/10.1145/2447976.2447990

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Proceedings of the First Workshop on Ethics in Natural Language Processing*, 53–59.

Telford, T. (2019, November 11). Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *Washington Post*. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

Tempelaar, D., Rienties, B., & Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS ONE*, *15*(6), e0233977. https://doi.org/10.1371/journal.pone.0233977

Tipton, E. (2014). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review*, *37*(2), 109–139. https://doi.org/10.1177/0193841X13516324

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *FairWare '18: Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101–120. https://doi.org/10.1177/0265532216679451

Waters, A., & Miikkulainen, R. (2014). GRADE: Machine Learning Support for Graduate Admissions. *AI Magazine*, *35*(1), 64. https://doi.org/10.1609/aimag.v35i1.2504

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving Retention: Predicting at-Risk Students by Analysing Clicking Behaviour in a Virtual Learning Environment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 145–149. https://doi.org/10.1145/2460296.2460324

Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., & Christopherson, R. M. (2010). The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS'10)* (pp. 327–337). Springer Berlin Heidelberg.

Wu, R., Xu, G., Chen, E., Liu, Q., & Ng, W. (2017). Knowledge or Gaming? Cognitive Modelling Based on Multiple-Attempt Response. *Proceedings of the 26th International Conference on World Wide Web Companion*, 321–329. https://doi.org/10.1145/3041021.3054156

Xia, M., Asano, Y., Williams, J. J., Qu, H., & Ma, X. (2020). Using Information Visualization to Promote Students' Reflection on "Gaming the System" in Online Learning. *Proceedings of the Seventh ACM Conference on Learning @ Scale*, 37–49. https://doi.org/10.1145/3386527.3405924

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should College Dropout Prediction Models Include Protected Attributes?. In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 91-100).

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 292–301.

Yudelson, M. V., Fancsali, S. E., Ritter, S., Berman, S. R., Nixon, T., & Joshi, A. (2014). Better Data Beat Big Data. *Proceedings of the 7th International Conference on Educational Data Mining*, 205–208.

Zhou, T., Sheng, H., & Howley, I. (2020). Assessing Post-hoc Explainability of the BKT Algorithm. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 407-413)