Algorithmic bias and discrimination through digitalization in education: a socio-technical view

Rebecca Eynon

University of Oxford, Oxford, UK, rebecca.eynon@oii.ox.ac.uk

Abstract

There is a long standing concern that the use of digital technologies in education risks maintaining, and in some cases exacerbating, existing social and educational inequity. Such debate has intensified due to the increasing use of AI in education that typically relies on advances in machine learning. There are multiple ways that AIEd can lead to unjust practices and outcomes. Employing a socio-technical perspective, this chapter focuses on the way that such systems encode certain biases and values that lead to discrimination of particular groups. It provides an overview of some of the main sources of bias that emerge when digitalizing education, from the outset of model creation to their application in practice; and outlines the strategies that can be made by data scientists to mitigate such bias. It then highlights the limitations of these numerical approaches, which largely ignore fundamental questions of justice and power. The chapter concludes by emphasising the importance of governance and regulation to mitigate discriminatory practices, calls for a more nuanced vision of how to best conceptualise and address social and educational inequity in current times, and the need for multi-stakeholder engagement to determine how, if at all, AI should be used in education.

1. Introduction

Scholars have long highlighted how education maintains and exacerbates social and educational inequity (e.g. Bourdieu and Passeron, 1990; Bernstein, 1996; Freire, 1974). Studies that have paid particular attention to the use of digital technologies in education have found similar trends. For example, digital inequality research has demonstrated how people who are less well connected are less able to engage fully with their education or to use digital means to support their learning interests across the life course (Robinson et al., 2018). Ethnographic work in schools has highlighted how schools in richer areas tend to use technology differently to schools in less well-off areas, continuing to facilitate forms of social sorting where different groups of students are supported on particular educational paths which lead to distinct life trajectories (Warschauer, 2004; Rafalow and Puckett, 2022). Others have demonstrated how open educational resources, designed as a way to open up educational and social opportunity, tend to promote particular ways of knowing and are most often used by those who already have the best access to education (Funes and Mackness, 2018; Adam, 2019). Such patterns continue beyond learning into the wider contexts of education. For example, digital technologies, often promoted as a way for young people and their families to access information about future educational opportunities to help address social inequity are also not neutral, with platforms designed to inform school choice via school reviews containing biased information – both in terms of who writes reviews and their content (Gillani et al., 2022). Such injustice, typically ignored by policy makers, were brought into sharp focus during the pandemic.

Of course, there have been some important instances, often using more participatory design practices and informed by critical pedagogy, where technology has addressed educational and social inequity (Garcia and Lee, 2020). However, it is reasonable to view these examples as the exception rather than the rule. Digital technology is not a straightforward fix for inequity. As Selwyn and Facer note, "regardless of technical sophistication or apparent ubiquity of adoption, technology use in education remains subject to and often reproductive of a range of persistent and pernicious inequalities" (Selwyn and Facer, 2021: 7).

Inequity in an era of Artificial Intelligence

This rather sobering picture has intensified with the current AI 'Spring'. As with each new wave of technological hype, the development of AIEd systems are promoted as a way to reduce educational inequity primarily through providing additional resources and 1-to-1 tutoring at scale (Holstein and Doroudi, 2021). Although resources are of course a central part of addressing questions of inequity, digital technologies are never straightforwardly "good things" that simply address an equity "gap" (Garcia and Lee, 2020). They are instead complex artefacts that embody social, cultural and political practices and values, that will be used and have different implications in different contexts. Thus, while there may be positives to the introduction of such systems into education, there are also likely to be negatives too, and these are not straightforward or uniform.

The use of AI in education is distinct from other forms of EdTech which do not use machine learning. In the majority of current AI systems, machines learn from past data to create a model that can be used on new data to make future predictions (Kizilcec and Lee, 2022). These predictions can be used to automate particular learning activities (e.g. determine which question or material the student should be introduced to next); inform teachers', students' or other educational stakeholders' future judgements (e.g. predicted grades to assist with college admissions, student dashboards to indicate progress relative to peers, or help teachers with determining areas of student misunderstanding or disengagement); or to take on tasks related to educational management and administration (e.g. proctoring, plagiarism detection or auto-transcribing lectures). Furthermore, this ever increasing automating of the teaching and learning process are increasingly not transparent to teachers, students and other stakeholders.

These shifts towards automation have led to concerns that such systems encode certain biases and values that lead to discrimination of particular groups. There have been concerns for example, that AI systems that predict future trajectories and/ or recommend particular learning paths may lead to a reduced learning experience for some students and limit future educational opportunities; reinforcing or exacerbating inequity along intersections of class, race and gender (Zeide, 2017). Others have highlighted how particular kinds of knowledge or ways of communicating may be favoured in systems designed to provide feedback on particular tasks such as automated essay scoring systems that favour white students (Dixon-Román et al., 2020). Others have raised alarms at how proctoring software favours students who behave in particular ways that are assumed to be "normal", that encode ableist assumptions; and require students with darker skin tones to shine lights directly on their faces while taking an exam to ensure the system can "see" their face (Madaio et al., 2022). Others highlight the questionable practice of systems that detect concentration and emotion in the classroom. Aside from important questions about the validity of measuring affective states (e.g. boredom, frustration, and engagement) from data generated from students writing or speech, facial expressions, gestures, and physiology there are also questions of bias (McStay, 2020). White students are again privileged, both because certain measures of physiology (such as oximeters) work better on fair skin; and because human coded data used to train facial recognition tends to positively favour the expressions of white westerners (McStay, 2020; Crawford, 2021). Beyond the context of formal education, these issues continue to exist in individuals' everyday experiences of AI, such as when searching for information or interacting with a digital assistant, where the experiences of white, well-off, well-educated males are typically privileged (Noble, 2018).

There are of course multiple questions that these brief examples generate. One key area is how best to theorise the complex interplay between AI systems and educational and social equity; and how they may cause harm due to a misallocation of opportunities and resources of various kinds, alongside representational harms where certain groups are portrayed in ways that diminish their identity (Mayfield et al., 2019; Madaio, et al., 2022). Here the focus is on one part of the issue, primarily on the implications of these systems in relation to discrimination. There is significant concern that people who are members of specific social groups (sometimes discussed as protected characteristics) are discriminated against through the use of these systems in educational systems which ultimately can favour certain groups of students over others.

Despite discrimination of any kind being illegal in large parts of the globe, there has been surprisingly little attention to this issue in the field of AI and education. While those working within educational data science recognize that it is critically important, in practice engagement has been limited. Very few studies within education that use machine learning actually use demographic data in any meaningful way to test issues of bias and discrimination (Paquette et al. 2020; Baker and Hawn, 2022). Indeed, very few of the EdTech tools on the market say anything about bias checks in their systems as part of their online advertising.

The ways that different 'knowledge traditions' (Furlong and Whitty, 2017) conceptualise questions of inequity, bias and discrimination in relation to AI and Education vary significantly (Eynon, 2023). In brief, more positivist and instrumental data scientists tend towards attempts at minimising or removing bias from particular systems while accepting the status quo; whereas those from more interpretivist backgrounds tend towards a focus on inequity in social structures, drawing attention to questions of justice and power. Taking a socio-technical perspective, this chapter aims to provide an overview of both perspectives of the debate by providing an overview of some of the main sources of bias that emerge across the 'ML pipeline' when digitalizing education, from the outset of model creation to the ways that systems are used for educational purposes; then discuss the challenges with such an approach; and some potential ways forward.

Bias across the 'ML pipeline'

AI systems are built via a series of practices, that can be described as the 'ML pipeline' (Suresh and Guttag, 2020). To understand how bias and discrimination emerges from such systems it is important to understand something about these practices.

Suresh and Guttag (2020) outline six steps in the 'ML pipeline'. First is defining the aims of the research and the creation of a suitable dataset to work with. This data set is often constructed from one or more pre-existing data sets, although can sometimes also involve purposively collected data for the specific research task. Second, the data set is split into at least two sets: the training data that is used to develop (i.e. train) the model and the test data, which as it names suggests tests the accuracy of the model on data the machine has not 'seen' before. Third, is the iterative development of the model using the training data to determine the model that performs the best on the task of interest. It typically requires a range of different activities, from selecting the algorithm, tweaking it, and using specific metrics to test the performance of the model. Fourth is evaluating the performance of the final model using the test data that has not been used before. At this stage, data scientists sometimes also use ‘benchmark data’ to compare the performance of the model relative to other existing models that have also been tested on the benchmark data. Fifth is where final decisions are made about how the model can be used in practice, and tweaking to enable this more practical use. This could include, for example, switching the outputs of a model that are expressed as probabilities into neat binary categories (e.g. the student will fail / will pass the course). Finally, the sixth stage of the pipeline is model deployment into ‘the wild’, i.e. the context and application to the actual population it is designed for. As noted above, this could be to provide teachers with information to help them with a decision or provide an automated decision (e.g. which course to recommend to a student or what test question to be asked next) (Suresh and Guttag, 2020).

Although this is of course a significant simplification of what is in reality a non-linear process, it demonstrates how each ‘stage’ requires a series of judgments and decisions that may be biased and lead to unfavourable social and educational outcomes (Suresh and Guttag, 2020). Indeed, a number of scholars have detailed how bias and potential for discrimination can emerge through measurement, problems with the model in how it performs for different groups, and how such models are applied in practice (Suresh and Guttag, 2020; Baker and Hawn, 2022; Kizilcec and Lee, 2022). These issues can lead to discrimination and a reinforcement of existing inequity in education or potentially introduce new problems and exacerbate existing injustice (Madaio et al., 2022).

The remainder of this section provides more detail about how discrimination can emerge in the creation and use of such systems, collapsing the six stages of the ML pipeline into three: measurement (i.e. the collection and curation of data to use for the model); model learning (i.e. the use of algorithms to create the model) and action (where the model is used for “real life” cases and judgements) (Kizilcec and Lee, 2022). It then considers questions of the potential biases of the data scientists themselves.

Measurement

Bias can occur during the process of defining the educational problem the model is designed to solve, and compiling an appropriate data set (Kizilcec and Lee, 2022). While both are important, most of the discussion has tended to try to ensure the measures used are as unbiased as possible. Three common types of bias at this stage of the ML pipeline are described as: historical, representation and measurement biases (Suresh and Guttag, 2020; Baker and Hawn, 2022).

Historical bias is the term used to recognise how ML models will always reflect existing inequity in education and society, regardless of the accuracy and validity of the sampling strategy. As such, these models serve to reinforce existing societal problems (Suresh and Guttag, 2020). For example, career guidance systems for young people built on existing data might not recommend to girls to consider a career in STEM, simply due to the very small proportion of women currently in STEM careers (Suresh and Guttag, 2020). Similarly, the popular development of models to predict grades that inform future education and career options for students are likely to reflect structural inequalities, for example, with those who are less well-off being predicted lower grades (as educational outcomes are often correlated with other measures of inequity) (Baker and Hawn, 2022; Kizilcec and Lee, 2022).

Representation bias is essentially a concern with ensuring that the data is representative of the students or other educational stakeholders the model is designed to speak to. These biases are caused due to problems in sampling (e.g. those who have limited digital access, or those from particular socio-demographic groups may be less well represented). If the data does not meaningfully represent the target group, then the model will be biased (Baker and Hawn, 2022; Kizilcec and Lee, 2022; Holstein and Doroudi, 2021). For example, substantial biases have been identified in facial analysis algorithms, with darker-skinned female faces being misclassified significantly more often than lighter-skinned males, due in large part to the under representation of darker skin tones in facial image data sets (Buolamwini and Gebru, 2018). Such errors happen even when minority groups in the population are sampled appropriately, as the final model may perform better for those in the majority group (Suresh and Guttag, 2020).

Under this first measurement phase, there are also a series of risks of bias related to the act of measurement itself. An important way this can occur is that the available data is not always a good proxy for the attributes it is designed to measure (Suresh and Guttag, 2020) – a point that is returned to below. Measurement problems can also emerge due to biases in the original collection of the data that are then re-used as part of the development of a new machine learning model. For example, if a data scientist needs to include a measure of dyslexia, existing administrative data that measures rates of dyslexia in student populations may be utilised. However, such data may be biased, as dyslexia is diagnosed at different rates depending on the socio economic background of the student (Baker and Hawn, 2022). Relatedly, assessment scores or measures of classroom behaviour may also be collected from the observations of experts (e.g. teachers or school administrators) and these measures may also reflect bias towards or against particular groups, that again become part of the models and therefore perpetuate bias (Holstein and Doroudi, 2021).

The politics and scale of data set creation

Discussions of bias in AI will likely feel familiar to any social science researcher. Yet, there is an important difference in how these data sets are constructed. The increasing focus on data intense approaches to machine learning that characterise the current AI 'Spring' has changed the way that databases for the development of ML models are created. Prior to the early 2000s, datasets were typically created for the key research goal. For example, if a data set of faces was required for face recognition then these were curated and labelled by a research team, where participants provided informed consent for their photos to be taken for this purpose[i]. Like all data sets, they were partial, imperfect and not representative, but were similar to many other kinds of research data that scientists collect (Crawford, 2021). As the demand for data increased, this changed. Researchers had to find significantly more data to support the machine learning process, and so began to collect "naturally occurring" data sets from everyday life. That is, using existing administrative data sets (e.g. facial images taken for law enforcement purposes) and the now ubiquitous practice of crawling data from the web at a scale previously unimaginable (Crawford, 2021).

Originally, data crawling on a mass scale was primarily the domain of the tech companies who kept the data for their own purposes. However, academic researchers and those working outside the tech industry also wanted access to these data sets, and so they got it by crawling the web and creating open source data sets for all to use. One example is the Common-Crawl dataset, created by a San Francisco based non-profit organisation that regularly crawls the web. The size of this and similar data sets is vast: as of April 2021, the Common-Crawl archive was around 320 TB in size and encompassed approximately 3.1 billion pages (Birhane et al., 2021:3).

Some of these data sets that are generated from crawling the web have been labelled by crowdworkers who have been recruited for the task. However, in the pursuit of ever bigger data sets, increasingly such data is not formally labelled and instead uses whatever associated information is available, such as the text used by the public when uploading or sharing images to the web. Aside from fundamental questions of informed consent, both approaches are likely to lead to significant problems, with the likelihood of bias and discrimination further intensifying in the second approach (Denton, 2021; Birhane et al., 2021; Crawford, 2021). Proposals for aiming to address bias in these data sets include: retractions of data sets or (perhaps more realistically) adding limits / licences to restrict use of the data set for academic purposes only and to not allow it to be used for models deployed in practice (Birhane et al., 2021). However, even from this short section, it is clear to see just how complex it is to meaningfully address measurement bias in this context (Birhane et al., 2021).

Model Learning

Model learning encompasses stages three and four in the ML pipeline outlined above. As part of this process, a number of decisions are made, including: which algorithm to use, how to deal with outliers, what tweaks might be needed to the parameters of the algorithm to improve the model fit and so on. All of this is carried out with a view to create the best model (as defined by accuracy and or efficiency) that can then be applied to a real life situation. All of these decisions are likely to have significant implications for the resulting model and may well lead to bias. This is because the data scientists engaged in the modelling process are likely to have specific assumptions about learning and education that are not made explicit, but could inform certain decisions when finessing the final model. However such practices are rarely reported, and the assumptions that have guided the decisions are rarely interrogated or made visible to anyone outside the core team (Kizilcec and Lee, 2022).

Towards the end of this phase, data scientists need to evaluate the final model, and this can include a component of checking for certain kinds of biases. This can include checking for aggregation bias which arises as a result of the use of one size fits all models which do not apply well to certain sub-groups. When aggregation bias occurs it can either lead to a model that tends not to be a good model for anyone in the population, or when representational bias is also present will be better for some than for others (Suresh and Guttag, 2020).

A second important check is for evaluation bias, in other words any biases in the model that occur due to problems with the representativeness of the test and / or benchmark data that is used to evaluate the performance of the model. As noted above, once the model has been developed, the model is tested on the unseen test data and sometimes also on benchmark data to compare the performance of the model against others than have been developed. This validation process is crucial for generalizability of the model to new contexts (Baker and Hawn, 2022). If the test data or benchmark data is not representative, then this causes problems. Indeed, the majority of ML models for education do not include information about the populations the models were tested on, making it difficult to assess evaluation bias (Paquette et al., 2020; Baker and Hawn, 2022). Beyond education, studies have also highlighted problems with benchmark data, which can contain undiscovered representation biases and that lead to underperformance of models for particular groups – and the resulting discrimination from the use of these models then also goes unrecognised (Buolamwini and Gebru, 2018; Suresh and Guttag, 2020).

Action

Action relates to the risks of bias and discrimination when the model is deployed into real world settings (Suresh and Guttag, 2020; Baker and Hawn, 2022). As noted above, different models are used as part of AI systems that are designed to action varied tasks. Some are used to automate a particular decision like who gets an A grade or who gets to go to a particular college. Others are used to inform human judgement, such as who is a good candidate for college entrance, or where a student's gap in knowledge is preventing them from getting an A. If it is a biased model, as a result of measurement and model learning, this will lead to discrimination and unequitable impacts. For example, under-predicting success for less well-off students may mean that they then do not get offered a college place or do not receive access to an intervention that would increase their chances of obtaining an A grade score.

Problems, or what is sometimes called "deployment bias", can also occur, "when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used" (Suresh and Guttag (2020:6). There are a number of ways a model can be used in such "off-label" ways. For example, if a teacher uses a system designed to detect boredom or frustration among students for formative purposes are used as part of a summative marking scheme (Baker and Hawn, 2022:1059).

There are other factors here too. For example, there could also be problems of the issue of interpretability, as in the case where complex models may be difficult to interpret and may not necessarily be that useful to educators or students in practice (Fischer et al., 2020). Indeed, one classic example of this is around the use of visualisation. Visualisation is commonly used as part of AI in education, in the creation of dashboards that are developed for a range of different users, from students, teachers and parents to policymakers. Yet, it is an imperfect art, and precisely how stakeholders use these, and indeed how these systems interact with their own individually held biases and beliefs is an important question (Holstein and Doroudi, 2021). Ultimately, there is a need for data scientists in education to really understand the contexts in which they are intervening and what they hope such models will achieve in practice (Eynon, 2023).

Expertise

In many discussions addressing bias in the ML pipeline, there is a sense that the biases and assumptions of the data scientists themselves can almost be removed from the equation. Yet, "knowledge production is never separate from the knowledge producer" (Boellstorff, 2013: n.p.). The framing of the problem to be solved, the selection of data, the creation of models from that data, and the uses to which these models are put, are all very much shaped by the expertise and perspectives of those who build them (Crawford, 2021). Data scientists are likely to be expert in data collection, management and analysis but are often from other fields such as engineering or physics, and have limited (if any) training or academic experience of education and social justice. This has a significant impact on a variety of stages of the production and application of data models.

Straightforward examples of the importance of world view include aspects of measurement bias which occur when the available data is not always a good proxy for the attributes it is designed to measure (Suresh and Guttag, 2020; Hershkovitz and Alexandron, 2020). For example, a correct answer does not straightforwardly equate to knowledge on a particular topic; and measurements of behaviour, such as downloading a file or playing a video, does not necessarily mean that the content has been read or watched. Relatedly, is a concern that a very narrow range of variables are used to capture a complex aspect of schooling. For example, college success is measured by a very small number of target variables (e.g. GPA) but this does not reflect the complexity of college success (Kizilcec and Lee, 2022; Eynon, 2022). Although often described as bias, these questions are also about broader questions of validity (e.g. McStay, 2020). There is a need to understand learning and education and to design systems that are appropriate for that context.

Beyond this, is the kind of taken for granted ways that such systems will inevitably be a "good thing" for education, yet it is important to ask who benefits the most from such systems. The majority of model learning models are developed and produced primarily by the most privileged groups in society, i.e. western, able bodied, white men. This risks models being produced that continue to favour a particular kind of world view, that is then applied to everyone, potentially perpetuating and reinforcing inequity. At the same time, such a privileged group are unlikely to be able to recognise questions of justice and power, despite the central role that they themselves, and the systems they create, play in maintaining the status quo (Birhane, 2021).

The limits of technical approaches to inequity

Alongside this removal (or at least recognition of) bias there remains fundamental normative questions about how inequity is conceptualised. Although these questions are actually present at all phases of the ML pipeline from measurement, model learning and action phases of the ML pipeline (Fischer et al., 2020, Hershkovitz and Alexandron, 2020), this has become most visible in the debates amongst data scientists around how to encode or quantify what "fairness" means when these models are used in practice.

For the most part, "fairness" in these discussions is relatively limited. It largely relates to the decisions around the predictions resulting from the model to try to ensure certain groups are not discriminated against (Kizilcec and Lee, 2022). When trying to ensure a system is fair, experts in machine learning select a mathematical model of fairness (e.g. that predictions are not conditional on group membership, or that the accuracy / performance of the model is equivalent across different groups). Then the ML pipeline is adjusted to ensure this concept of fairness can be met. Data scientists can make adjustments at the stage of collating and refining the database, as part of the modelling process, or when evaluating and adjusting the final model in the post-processing stages (Barocas et al., 2019; Suresh and Guttag, 2020).

Attempting to reduce bias and discrimination in such systems is, without doubt, essential. However, it is not without its challenges. There is growing recognition amongst data scientists that there is no one correct measure of fairness, and that the ways to address it very much depend on the particular case (Suresh and Guttag, 2020). Scholars also argue that it is important that the choice of fairness metric is justified and is chosen in way that encourages debate amongst relevant stakeholders (Kizilcec and Lee, 2022).

Such concerns about bias and farness are closely related to calls to improve the transparency of systems. For example, through the use of interpretable machine learning to enhance the explainability of such systems (Kizilcec and Lee, 2022; Holstein and Doroudi, 2021); and the need for third party checks on EdTech systems with respect to validity and bias prior to being rolled out to classrooms (Boninger et al. 2017; Regan and Jesse, 2019). Relatedly, there have been calls to ensure AI systems in Education are accountable both in terms of who to hold to account, and what they can be accountable for (Hakimi et al., 2021).

While all of these areas of focus are important it points to the wider set of problems of trying to operationalise fairness. The approach to addressing bias and making models fair, neglects wider questions of equity and justice in relation to the use of these systems. As Green and Hu note, "Fair machine learning is hard not because of statistical or computational challenges, but because striving for fairness is ultimately a process of continual social negotiation and adjudication between competing needs and visions of the good" (Green and Hu, 2018: 3).

This "social negotiation" is largely absent from many discussions of algorithmic fairness by data science in education and beyond (Paquette et al., 2019). The focus on the data and the models tends to divorce these systems from the wider social structure of which they are part (Denton, 2021; Crawford, 2021). It locates issues of bias and discrimination at the level of the model, or a particular instance as opposed to “interrogating systemic discrimination”; whether wider society is fair is never questioned (Green and Hu, 2018:3). As Dave argues, there is a need to change the language, away from a focus on the flaws of an individual model of an AI system, towards a recognition of injustice in society (Dave, 2019). In part, this is facilitated by the focus on discrimination, which has a long history of presenting the problem as one of individual “bad actors” rather than structural issues, where the removal of the bad actor is possible and the status quo remains undisturbed (Hoffmann, 2019).

Ultimately, the vast majority of work in machine learning tends to view data as an objective “thing” that can provide a window on the world (Eynon, 2023). It ignores how these systems are part of a socio-technical social structure that reinforces and reconfigures inequity (Winters et al., 2020). Indeed, such systems can actively work against positive change, as using historical data to predict future patterns can entrench existing patterns of inequity (Green and Hu, 2018). As Hoffman argues, it is important to fully conceptualise how “humans and technology co-conspire to not just passively reproduce but actively uphold and reproduce discriminatory social structures” (Hoffmann, 2019:905). It also enables ML researchers not to take responsibility for wider harms from the models they build, and their role (consciously or not) in actively reproducing inequity in ways that continue to work in their favour (Hoffmann, 2019).

In addition, such a computational and statistical agenda tends towards a one dimensional framing of discrimination and an over focus on allocation issues as the primary focus of inequity (Hoffmann, 2019). If discrimination is considered at all in AI for education it tends towards a focus on understanding what the bias in a model might mean for one particular group (e.g. Black vs White or Rich vs Poor students)[ii]. However, this leads to problems in two respects. First, it ignores the crucial intersectional basis of discrimination (Crenshaw, 1989) and tends to only consider groups in relation to disadvantage as opposed to the maintenance and reproduction of privilege (Hoffmann, 2019). Further, the common focus on allocation (or misallocation) or resources and opportunities, while important, tends to ignore other ways that injustices arise within our education system. For example, such injustices might occur due to the content of curriculum or the design of pedagogical approaches that are not sensitive to varied educational contexts; or due to the extractive nature of the use of educational data, which tends to objectify and exclude specific groups (Mayfield et al., 2019; Regan and Jesse, 2019; Madaio, et al., 2022).

Resetting the agenda

These limits are recognised by many of those working on questions of algorithmic fairness, a field that has emerged in response to concerns about AI and inequity. Some of these debates relate to questions of objectivity, validity, reliability and the reduction of bias in algorithmic systems; yet others relate to wider questions of how to conceptualise questions of equity and power which cannot be adequately addressed through particular quantitative practices, however sophisticated (Hanna et al., 2020). These include a growing number of conferences, academic papers and workshops to address questions of bias and discrimination (e.g. ACM FAccT, workshops on ethics at AIED); numerous AI codes and frameworks produced by an array of governments, corporations, and the third-sector (Schiff, 2022); institutional ethical codes of practice that tend to focus on the related issues of security, transparency, accountability, trust and reliability (Hakimi et al., 2021); and moves to improve regulatory powers.

Research in this area specifically within the context of education lags behind debates in other areas such as law and health. This is somewhat surprising as it is in direct contradiction with the social values of many data scientists in this space who wish to improve education in some way (Paquette et al., 2020). This is compounded by additional problems also experienced beyond education, not least the increasing power of the private sector in this space, both in developing AIED and in determining governance and regulation of AI (Williamson, et al., 2023); and legal frameworks that are not keeping up with the social, technical and cultural implications of AI and associated approaches (Berendt et al., 2020; Holmes, 2022). At present, discussions of what is just or equitable in education tend to come from beyond the data science community, with limited interaction between more sociological and data centric communities, although there are moves in that direction (Holmes, 2023).

It is clear then, that there is far more to do. Issues of governance and in particular regulation remain central for mitigating discriminatory practices. Yet this needs to be connected to the larger, more complex questions of how to best conceptualise and address social and educational inequity in societies where AI is becoming part of the social fabric. A central part of this is "to interrogate and rethink what equity means for educational AI" (Madaio., et al., 2022:225). A central approach is to enable more participatory ways to create AI systems in education that stakeholders can agree are desirable, or indeed to decide not to have them at all. In order to achieve this multiple approaches and conceptual perspectives are required. Such work may include: changing data science curricular in order to bring into focus the highly political nature of data science in education and enable more awareness of questions of equity in future design (Green, 2021), to facilitate more meaningful links between sociologists of education and educational philosophers with the AIED community to develop better conceptual framings on inequity (Williamson and Eynon, 2020), and to promote the use of more participatory, emancipatory and inclusive approaches to design where teachers, students and other stakeholders are meaningfully included (D'Ignazio and Klein, 2019; Madiao et al., 2022; Winters et al., 2020). Of course, it is important not to see participatory projects in all their forms as a straightforward panacea (Madaio et al., 2022). Nevertheless, they are a central way to make visible and change the inequitable structures in society.

Conclusion

There is significant concern that the digitalization of education risks maintaining, and may be exacerbating, existing social and educational inequity. Systems designed to: facilitate school choice, evaluate essay writing, judge the effectiveness of teachers, and assess students are just some of a growing number of examples of systems that can favour certain groups over others.

Such concerns grow ever greater as AIED scales. No one system can be considered in isolation, and it is likely that earlier biases and risks of discrimination are compounded as new models are developed on the basis of past actions. However, the precise scale of the problem is largely unknown given the lack of computational studies within education that even test issues of bias and the absence of any engagement with this issue from the vast majority of EdTech products used across the world.

The calls to identify and remove biases from these systems and try to make them fair is an important endeavour. Yet numerical models can only go so far, not just because there are likely practical limits on the extent to which bias can be mitigated in AIED, but also because they reframe a complex social question of equity into a numerical judgement and an individualistic reponse. Such bias and discrimination needs to be understood in a relational sense, that goes beyond any one individual model to a focus on questions of the interplay between AI and social structure, and draw on theories from more critical literatures.

Alongside developments in governance and regulation to mitigate discriminatory practices, there is a need for a more nuanced vision of how to best conceptualise and address social and educational inequity in current times, and the need to determine how, if at all, AI should be used in education. Such debates and visions can only be determined through the participatory actions from an array of stakeholders, including the general public, students, teachers, policy makers, Edtech developers and academic communities to create equitable futures for all.

References

Adam, T. (2019). Digital neocolonialism and massive open online courses (MOOCs): colonial pasts and neoliberal futures, *Learning, Media and Technology*, 44(3), pp.365-380.

Baker, R.S. and Hawn, A. (2022). Algorithmic bias in education, *International Journal of Artificial Intelligence in Education*, 32(4), pp.1052-1092.

Barocas, S., M. Hardt, and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. https://fairmlbook.org.

Berendt, B., Littlejohn, A. and Blakemore, M. (2020). AI in education: learner choice and fundamental rights, *Learning, Media and Technology*, 45(3), pp.312-324.

Boellstorff, T. (2013). Making big data, in theory, *First Monday*, 18(10).

Boninger, F., Molnar, A., and Murray, K. (2017). Asleep at the Switch: Schoolhouse commercialism, student privacy, and the failure of policymaking. National Education Policy Center. Available at. https://nepc.colorado.edu/publication/schoolhouse-commercialism-2017 (Accessed: 5 January 2023)

Bourdieu, P. and Passeron, J.C. (1990). *Reproduction in education, society and culture*. 2nd edn. London: Sage.

Bernstein, B. (2000). *Pedagogy, symbolic control, and identity: Theory, research, critique*. Oxford: Rowman & Littlefield.

Birhane, A., Prabhu, V.U. and Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. https://arxiv.org/abs/2110.01963.

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach, *Patterns*, 2(2).

Buolamwini, J., and Gebru. T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency*, New York, 23-24 February. https://proceedings.mlr.press/v81/buolamwini18a.html.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Denton, E., Hanna, A., Amironesei, R., Smart, A. and Nicole, H. (2021) On the genealogy of machine learning datasets: A critical history of imagenet, *Big Data & Society*, 8(2).

Dixon-Román, E., Nichols, T.P. and Nyame-Mensah, A. (2020). The racializing forces of/in AI educational technologies, *Learning, Media and Technology*, 45(3), pp.236-250.

D'Ignazio, C. and Klein, L. (2019). *Data Feminism*. Cambridge: MIT Press.

Eynon, R. (2022). Datafication and the role of schooling. In, Pangrazio, L., and Sefton-Green, J. (Eds). *Learning to Live with Datafication: Educational Case Studies and Initiatives from Across the World*. London: Routledge.

Facer, K. and Selwyn, N. (2021). Digital Technology and the Futures of Education: Towards 'Non-Stupid' Optimism. Paris: UNESCO. Available at, https://unesdoc.unesco.org/ark:/48223/pf0000377071 (Accessed: 5 January 2023)

Fischer, C., Parados, Z. and Baker, R. (2020). Mining Big Data in Education: Affordances and Challenges, *Review of Research in Education*, 44(1), pp.130-160.

Freire, P. (1970). *Pedagogy of the oppressed*. New York: Continuum books.

Funes, M. and Mackness, J. (2018). When inclusion excludes: a counter narrative of open online education, *Learning, Media and Technology*, 43(2), pp.119-138.

Garcia, A. and Lee, C.H. (2020). Equity-centered approaches to educational technology. Handbook of research in educational communications and technology, pp.247-261.

Gillani, N., Chu, E., Beeferman, D., Eynon, R. and Roy, D. (2021). Parents' Online School Reviews Reflect Several Racial and Socioeconomic Disparities in K–12 Education, *AERA Open*, 7.

Green, B. (2021). Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*, 2(3), pp. 249-265.

Green, B. and Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning. https://www.benzevgreen.com/wpcontent/uploads/2019/02/18-icmldebates.pdf

Hakimi L., Eynon R., and Murphy, V. (2021). The ethics of using digital trace data in education: A thematic review of the research landscape, *Review of Educational Research*. 91(5), pp.671-717.

Hanna, A., Denton, E., Smart, A. and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness, Proceedings of the 2020 conference on fairness, accountability, and transparency, Barcelona, January 27 – 30. https://arxiv.org/abs/1912.03593

Holstein, K. and Doroudi, S. (2021). Equity and Artificial Intelligence in Education: Will AIEd Amplify or Alleviate Inequities in Education? https://doi.org/10.48550/arXiv.2104.12920

Hershkovitz, A. and Alexandron, G. (2020). Understanding the potential and challenges of Big Data in schools and education, *Tendencias pedagógicas*, 35, pp.7-17.

Hoffmann, A.L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, *Information, Communication & Society*, 22(7), 900-915.

Kizilcec, R. F. and Lee, H. (2022). Algorithmic fairness in education. In, W. Holmes & K. Porayska-Pomsta (Eds.), *Ethics in Artificial Intelligence in Education: Current Challenges, Practices and Debates*. London: Routledge.

Madaio, M., Blodgett, S.L., Mayfield, E. and Dixon-Román, E. (2022). Beyond "Fairness": Structural (In) justice Lenses on AI for Education. In, W. Holmes & K. Porayska-Pomsta (Eds.), *Ethics in Artificial Intelligence in Education: Current Challenges, Practices and Debates*. London: Routledge.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E. and 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, Florence, August. https://aclanthology.org/volumes/W19-44/

McStay, A. (2020). Emotional AI and EdTech: serving the public good? *Learning, Media and Technology*, 45(3), pp.270-283.

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

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A. and Baker, R. (2020). Who's Learning? Using Demographics in EDM Research, *Journal of Educational Data Mining*, 12(3), pp.1-30.

Rafalow, M.H. and Puckett, C. (2022). Sorting Machines: Digital Technology and Categorical Inequality in Education, *Educational Researcher*, 51(4), pp.274-278.

Regan, P.M. and Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking, *Ethics and Information Technology*, 21(3), pp.167-179.

Robinson, L., Chen, W., Schulz, J. and Khilnani, A. (2018). Digital Inequality Across Major Life Realms, *American Behavioral Scientist*, 62(9), 1159–1166.

Schiff, D. (2022). Education for AI, not AI for Education: the role of education and ethics in national AI policy strategies, *International Journal of Artificial Intelligence in Education*, 32(3), pp.527-563.

Suresh, H. and Guttag, J.V. (2020). A framework for understanding unintended consequences of machine learning. https://arxiv.org/abs/1901.10002v3.

Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide*. Cambridge: MIT press.

Williamson, B. and Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education, *Learning, Media and Technology*, 45(3), pp.223-235.

Williamson, B., Eynon, R., Knox, J., and Davies, H. (in press). 'Critical perspectives on AI in education: Political economy, discrimination, commercialisation, governance and ethics.' In, Du Boulay, B., Mitrovic, T., and Yacef, K. (Eds). *The Handbook of Artificial Intelligence in Education*. London: Routledge.

Winters, N., Eynon, R., Geniets, A., Robson, J. and Kahn, K. (2020). Can we avoid digital structural violence in future learning systems? *Learning, Media and Technology*, 45(1), pp.17-30.

Zeide, E. (2017). The structural consequences of big data-driven education, *Big Data*, 5(2), pp.164-172.

[i] See Crawford (2021) for a discussion of the case of the Face Recognition Technology (FERET) program in the US as an example of this where participants gave consent for their images to be used, and the data was carefully curated and labelled by researchers.

[ii] See also https://www.pcla.wiki/index.php/Algorithmic_Bias_in_Education. This wiki provides a very helpful and much needed overview of studies of algorithmic bias in education, but highlights the typical one dimensional view of many studies.