Design propositions for the critical analysis of educational machine learning-based applications using emancipatory pedagogy

Bruno Elias Penteado

Instituto de Ciências Matemática e Computação – Universidade de São Paulo (USP)
13566-590 – São Carlos – SP – Brasil

brunopenteado@alumni.usp.br

Abstract. *In education, machine learning applications provide support and analytics insights to students, teachers, and administrators. However, not all of us are treated equally by these technologies. The algorithmic bias may reflect unequal opportunities for individuals based only on their demographic data. Following a design science research approach, we investigate multiple sources of bias in the machine learning pipeline and use emancipatory pedagogy as kernel theory to elaborate design propositions to mitigate this problem. We correlate the sources of bias with potential actions, providing theoretical lenses to handle bias throughout the development of intelligent educational systems. These principles should provide researchers with a critical analysis of the development of intelligent systems in education.*

Resumo. *Na educação, aplicações baseadas em aprendizado de máquina fornecem suporte e insights analíticos a alunos, professores e administradores. No entanto, nem todos nós somos tratados igualmente por essas tecnologias. O viés algorítmico pode refletir em oportunidades desiguais para os indivíduos com base apenas em seus dados demográficos. Seguindo uma abordagem de pesquisa em design science, investigamos múltiplas fontes de viés no pipeline de aprendizado de máquina e usamos a pedagogia emancipatória como teoria kernel para elaborar propostas de design para mitigar esse problema. Correlacionamos as fontes de viés com ações em potencial, fornecendo lentes teóricas para lidar com o viés no desenvolvimento de sistemas educacionais inteligentes. Esses princípios devem fornecer aos pesquisadores uma análise crítica do desenvolvimento de sistemas inteligentes na educação.*

1. Introduction

Machine learning (ML) systems have brought a new wave of innovation to different industries in the last decade, including educational technology. Computer programs built

DOI: 10.5753/sbie.2022.225710

from training data rather than handcrafted code excel at tasks previously thought difficult (Kane et al., 2021), such as image and speech recognition, autonomous vehicles, language translation, conversational agents, etc. Moreover, these algorithms are often viewed as inherently fair and objective (Lee, 2018). Educational technologies increasingly use digital data and predictive models to provide support and analytic insights to students, instructors, and administrators (Kizilcec & Lee, 2022). Tutoring systems, adaptive systems, automated scoring systems, adaptive testing, dropout prediction, and graduate admission - henceforth called educational ML-based applications - are some of the applications that use machine learning techniques as their core approach. These predictive models may be presented in different forms to the users. For example, systems can be designed to embed the predictions overtly, such as automatic detectors of affect or help-seeking behavior, in features of tutoring systems or learning management systems to offer automatic support. On the other hand, the predictive models may be presented in dashboards to raise awareness of tasks subject to human analysis, such as monitoring students' progress or topic mastery.

This vision of algorithmic objectiveness has been recently challenged since humans design algorithms and, as such, cannot be considered neutral, often encoding different biases of their developers or the society as a whole (Baker & Hawn, 2021). The problem of justice and fairness in education has a long history. Education large-scale testing has faced bias since the 1960s and laid out many aspects of the modern literature on algorithmic bias and fairness (Hutchinson & Mitchell, 2019). With the adoption of computers and predictive technology - sometimes in high-stakes applications, such as dropout prediction, automatic essay scoring, graduate admissions, and knowledge inference - it is replicated in scale, often without proper critical analysis of the effects on a population. Even the fields of Educational Data Mining and Learning Analytics emerged in the late 2000s as a form to congregate researchers around this topic. Some researchers have pointed out the possible uneven effectiveness and lack of generalizability across populations in educational algorithms (c.f. Bridgeman et al., 2009; Ocumpaugh et al., 2014).

Kizilcec & Lee (2022) frame this as a fairness problem in education, focusing on concerns of bias and discrimination. Other authors favor the use of *unfair* over *biased*, preserving bias for its statistical meaning and using fair/unfair for its social/moral implications (Baker & Hawn, 2021). The term *algorithmic bias* is often used as a term to define, for example, inequitable predictions across identity groups (Gardner et al., 2019), unwanted or societally unfavorable outcomes (Suresh & Guttag, 2021), or an algorithm whose decisions are skewed towards a particular group of people (Mehrabi et al., 2019).

Kizilcec & Lee (2002) state that although it is not a technological problem, we, as technology researchers, must account for that: "*it is unfair if students from low-income families score lower test scores for lack of access to study resources available to high-income families, but it is especially unfair if they score lower because their teacher - or an algorithmic scoring system - is biased against them.*" Algorithmic systems in

education can entrench historical inequities while obscuring the root cause and amplifying the effect [Mayfield et al., 2019; Shum, 2018]. Research that explores unfairness in these systems tends to conceptualize this issue as a purely mathematical or engineering problem, often avoiding the needed investigation into the set of values and systems of power that shapes them [Karumbaiah & Brooks, 2021; Gardner et al., 2019]. Most of the research on empirical algorithmic bias in education examined three categories: race/ethnicity, nationality, and gender (Baker & Hawn, 2021).

Karumbaiah (2021) points out that choosing theories, design elements, and methods may embed unintended bias. For instance, some theories were built from a specific societal stratum (e.g., WEIRD - white, educated, industrial, rich, and democratic) and caused replication issues on empirical data not collected from these same settings. In design, the presentation of features or field inputs may mask confounders from data - e.g., by not including non-binary gender data for user forms. Data annotation is a task crucial for ML and often handcrafted by humans; these humans may also disagree on standard subjective states in education (e.g., affect) and may be different in diverse cultures.

These elements configure a power imbalance, where dominant groups legitimate their oppression as a natural thing. For instance, it is aligned with the argument of Karumbaiah & Brooks (2021), which states that coloniality (often assumed as a thing from the past) continues to shape technological advancements in education. As automated and data-driven systems are becoming more widely implemented in classrooms, online settings, testing, admission and hiring decisions, school security, and more, it has become necessary to critically examine the principles that underlie these systems [Karumbaiah & Brooks, 2021; Lee & Kizilcec, 2020; Blikstein, 2018].

In this paper, we review how the literature conceptualizes algorithmic bias and its sources in education and frame this problem under the lens of emancipatory pedagogy (EP), suggesting some design propositions to help mitigate this issue in educational technology. Section 2 presents possible sources for algorithmic bias and the theoretical framework to analyze this problem. Section 3 reviews the literature on unfairness in education and some forms of mitigation already proposed. Section 4 describes the methodological procedures of this work. Next, section 5 shows the resulting design propositions and discusses how they can be applied in educational machine learning-based technology.

2. Theoretical background

In this section, we detail how the literature has approached the sources of bias in the machine learning pipeline to understand where possible distortions may be introduced. Next, we detail the aspects of the kernel theory considered to derive mitigation techniques for human emancipation.

a. Sources of algorithmic bias

In Suresh & Guttag (2021), the authors list seven sources of the potential introduction of bias in the machine learning pipeline. Algorithm bias here is considered a possible source of harm throughout the ML process that can lead to societally unfavorable outcomes in specific student subpopulations (Suresh & Guttag, 2021) but also in cases where a model's predictive performance unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class (Mitchell et al., 2021). A comprehensive comparison of sources can be found in Baker & Hawn (2021)

- *Historical bias*: even if it reflects the world accurately, it can still reflect harm to a population, such as reinforcing a stereotype;
- *Representation bias*: the sampling underrepresents some part of the population, failing to generalize well;
- *Measurement bias*: choosing, collecting, or computing features and labels to use in predictions by using problematic proxies for some constructs which are poor reflections or generated differently across groups
- *Aggregation bias*: the use of a one-size-fits-all model for data where subgroups should be considered differently - resulting in a model that is not optimal for any group or fits the dominant population
- *Learning bias*: modeling choices may amplify performance disparities across different examples in data when the prioritization of one objective may damage another one;
- *Evaluation bias*: the model is built on training data that does not represent the target population, with a desire to compare different algorithms regardless of their importance;
- *Deployment bias*: is the mismatch between the problem a model intends to solve and the way it is used.

This common approach is considered a *downstream* approach to bias awareness. Its focus has mainly been on investigating bias in predictive modeling, particularly its downstream stages like model development and evaluation. Karumbaiah et al. (2021), on the other hand, argue that *upstream* sources (i.e., theory, design, training data collection method) also contribute to the bias in these systems, highlighting the need for a nuanced approach to researching fairness.

- *Theory*: it drives the conceptualization, the proper methodology, data collection, algorithmic model, and design choices in research. However, based on restricted (and biased) settings, theory building may fail to generalize to other settings. For instance, the authors challenge the generalizability of a widely accepted model for affect detection in tutoring systems, presenting evidence of non-conformance of empirical data with students from varied nationalities;

- *Design*: user inputs, interface presentation, content sequencing, visual aids, or even the language used for human interaction. For instance, demographic forms may not allow filling non-binary gender information or the use of female pedagogic agents favoring representation and engagement of female students;
- *Method*: the data collection, labeling, algorithm model selection, relevant metrics, construct operationalization, and other subjective choices. For instance, the observation of off-task behavior - widely studied in tutoring systems research - is highly contextual and subjective, with positive self-regulation actions (e.g., taking a break) may be confused with disruptive behaviors (e.g., playing or wondering);

b. Emancipatory pedagogy

Emancipatory pedagogy is considered a critical theory, a research paradigm that believes that reality is interpreted or constructed by social actors as individuals or in social groups. It is mainly concerned with issues of power and justice and the ways that the economy, matters of race, class, gender, ideologies, discourses, education, religion and other social institutions, and cultural dynamics interact to construct a social system (Williamson & Johnson, 2018). We argue that imposing the oppressor's reality is a form of bias reinforcement for the minority groups, which can hardly cope with this context. For instance, considering design features as ‘normal’ in cultural contexts where they do not make much sense or reinforce historical biases.

Freire's seminal Pedagogy of the Oppressed (1970) described it as a "pedagogy of people engaged in the fight for their liberation" in the context of a postcolonial Brazil. Freire described how oppressed groups could achieve emancipation by gaining and promoting awareness of their reality and taking ownership of their struggle. His work advocated for emancipation through pedagogy in postcolonial contexts, with the oppressed gaining and promoting awareness of their reality and taking ownership of their struggle. For Freire, education should be a ‘practice of freedom’ with the potential to transform rather than conform (Freire, 1970). Young (2017) identified four functions in the emancipatory pedagogical process: awareness of the oppressed, problem awareness, system awareness, and solution enablement. In this work, we sought to go through these phases, from contextualizating algorithmic fairness (awareness of the oppressed) to developing design propositions (solution enablement) - although not detailed here. Kane et al. (2020) elicit four universal themes of emancipatory pedagogy for modifications in an ML model: i) *humanization*, where the user is not considered as an object and his/her humanity is considered during the optimization process; ii) *human experiences*, contextualizing the users’ experiences to learn about the world and their place; iii) *communication* between ML systems and users, to overcome the opacity of decision-making criteria, specific for the system-user partnership; iv) the balance of *freedom and authority*, the need for emergent rules and accountability while involving freedom of thought, action, and belonging.

3. Related work

Previous studies in education literature have investigated inequalities and inequities in educational opportunities and outcomes, such as school segregation and achievement gaps. However, since the Coleman Report in 1966 (Coleman, 1966), academic achievement gaps have become a focus of educational reform efforts, arguing that a combination of home, community, and in-school factors give rise to systematic differences in educational performances between groups of students based on their socioeconomic status, race-ethnicity, and gender (Kao & Thompson, 2003).

Research outside education categorized harms caused by algorithmic bias into allocative and representational (Crawford, 2017). *Allocative* harms are related to withholding opportunities or resources from specific groups or the unfair distribution of a good across groups; *representational* harm is the systematic representation of some group in a negative light or a lack of positive representation. Recent research has focused on the sources of bias, as detailed in the previous section. *Downstream* and *upstream* sources of bias have been investigated, aiming to detect, evaluate, and mitigate potential biases as soon as they occur in the ML pipeline. For educational applications, the decision-making - automated or as support for human judgment - is highly dependent on these concepts.

Much recent work addressing algorithmic bias has focused on mitigation at downstream sources - model evaluation and the ML pipeline postprocessing steps. Kizilcec & Lee (2022) also explores some mitigation tasks, mainly for downstream steps (measurement, model learning, and output presentation), highly concentrated on the statistical and engineering lenses of bias detection and correction. Some resources to aid can be found as open source, for instance, IBM's *AI Fairness 360* (https://aif360.mybluemix.net), with tools to examine, report, and mitigate some forms of algorithmic bias during model development, evaluation, and deployment. Suresh & Guttag (2021) propose some general forms of mitigation for the multiple sources of bias but in a simplistic manner without acknowledging the root causes. Baker & Hawn (2021) encourage an upstream approach with participatory design and evaluation and provide four recommendations: improving data collection, improving tools and resources, creating a structure to encourage openness, and broadening the community.

In addition, critical theories have been used in information systems research; however not as popular as positivist and interpretive paradigms. For example, considering emancipatory pedagogy, Kane et al. (2020) propose a design theory for emancipatory assistants to engage with human users to help them understand and enact, considering an oppressive future of ubiquitous monitoring and behavior control. Young (2017) also uses this theory to analyze how native Americans appropriated ICT tools to foster their cultural identity restoration.

Our work extends this literature by applying the theoretical lenses of critical theory to analyze the sources of algorithmic bias and the underlying oppression, not

limited by statistical or technological mitigation efforts. We posit, in line with Kane et al. (2020), that this theoretical framework is appropriate for our context, considering that, as much as in postcolonial Brazil, the subjugated populations do not know how to rectify the oppression they face, the same way as these populations must become aware of algorithmic unfairness against them and how to change it. This kernel theory helps in that direction, supporting the creation of design propositions and practices.

4. Methodology

We adopted a design science research (DSR) approach to this problem to enable the creation and use of artifacts to intervene in a given situation. DSR is an essential paradigm in information systems research, combining theoretical and technical with rigor and relevance. Its research focuses on theory building and knowledge about how an IT artifact behaves as it does (Gregor & Jones, 2007) - the latter includes both the system and the system-in-use by the human. The relevance is evaluated from the impacts of real-world problems. The scientific rigor for building IT artifacts is derived from kernel theories. A *kernel theory* is based on natural or social sciences, which provides insight into how to solve a problem and understand what could be done through design to achieve a particular solution. A possible outcome in DSR research is the derivation of design propositions - prescriptive knowledge (the core of DSR) using the existing published research base, offering a general template for creating solutions for a particular class of field problems (Denyer et al., 2008).

To define the problem, we used the upstream sources of bias, considering its effects on educational ML-based applications. Some authors have argued that moving upstream may mitigate downstream problems (Baker & Hawn, 2021), but this is an open-ended research question. Using emancipatory pedagogy (Freire, 1970) as a kernel theory, we provide design propositions toward the awareness of the oppressed subpopulations in algorithmic biases, aiming for human emancipation. We build upon the four principles from Kane et al. (2020) since it operationalized the theory in the context of the ML pipeline. We find these approaches suitable by considering machine learning bias as a form of naturalization of oppression for underrepresented groups of people and emancipation to break this out. As an evaluation, we provide illustrative narrative examples of the propositions and their effects from the literature.

5. Results

By combining the four principles of emancipatory pedagogy, we derive the following design propositions in Table 1, combining the principles of EP and the sources of algorithmic bias.

Table 1. Design propositions based on emancipatory pedagogy principles and upstream sources of bias.

Propositions	EP Principles	Description	Upstream source

DP1 - Design for a humanized perspective	M1 - humanization M2 - human experiences	Individuals' interests are not alienated from the learning process, and the metrics of the system are designed for pedagogical purposes. Systems should adapt according to the users' living experiences and contexts, providing equity in opportunities.	*Theory*: build on the acknowledgment that behavior is affected by factors such as culture and institutions *Design*: adapt features and language according to the user context *Method*: select proper educational metrics and test for the presence of algorithmic biases.
DP2 - Design for human agency	M3- communication M4 - freedom & authority	Explainable decisions should be clear to the user when possible to raise awareness on why the system behaves like that (adaptation, recommendation, etc.). This feature should include an option for opting out of adaptive features or data when the user feels like it.	*Theory*: explain why those attributes are relevant to the problem; *Design*: provide users with an explanation of the decision-making based on their attributes and the possibility of opting out of adaptive features. *Method:* choice of explainable algorithms

Design proposition 1 (DP1) builds upon a humanized perspective on the relationship between users and the computer system. The main requirement is avoiding treating users as objects (e.g., a cog in a machine), in the sense that the oppression is continuously relaxed and the user treated with dignity and adapted to their contexts (task- and demographic-wise). A form of dehumanization is the uncritical use of non-pedagogical metrics to evaluate these systems. Bachmair et al. (2018), based on cultural studies on metrics, argue that metrics should be pluralistic modes of representation, offering opportunities to emancipate the learning process from policies that merely treat "policy subjects as objects of intervention needing remediation" (Gulson & Webb, 2018). Inherent within the notion of merely exploiting learning analytics to 'fix' learners according to normative educational goals and cultural practices is a high risk of alienation and exclusion as the education machine fails to respond to the "linguistic, literate, and cultural pluralism" that is central to the democratic project of schooling. They propose two categories of metrics to counter dynamic to the alienation of monitored learners and assessed learning outcomes: the awareness of our everyday life and narrative interpretation. Standard metrics such as time on task, engagement level, and feedback effectiveness are positivist metrics that should be favored in contrast to classical ML metrics, such as accuracy or precision-recall curves. Even for the cases of classical metrics, they should weigh the classification error costs in specificity and sensitivity - such as the harm in labeling a student a potential at-risk when it is not the case. Emergent metrics for bias detection should also be used, and existing tools and processes can be

found for this matter (Boza & Evgeniou, 2021). In addition, in Freire's conception, education cannot just involve abstract ideas but contextualize learners' experiences to learn about the world. Design choices, such as the gender of pedagogical agents or the language applied, can make a big difference. For instance, in the study of Finkelstein et al. (2013), students showed a better performance in science when the adaptive system used a similar dialect to their native tongue (African American Vernacular English). Dialectal differences help explain for the systematically reduced test scores of children of color compared to their Euro-American peers. Ogan et al. (2012) showed how Latin Americans appropriated tutoring systems differently than expected. These systems were developed and optimized for individual use but used collectively in other countries' classrooms - for cultural or poor infrastructural reasons.

Design proposition 2 (DP2) expands on human agency - i.e., the capacity for human beings to make choices and to impose those choices on the world. The main requirement is to raise awareness if a decision is unfair based on the demographic attributes of the user and allow them to act upon it. When requested, an ML-based application should inform users of its rationale for decision-making, presenting to users how different attributes contributed to the decision. Khosravi et al. (2022) identified specific educational needs in explainable AI (XAI), with gains such as higher trust and awareness for learning decisions. However, these explanations must be carefully designed so that users can appropriate them. Francis et al. (2020) present papers showing that learning analytics can improve student agency and enable greater personalization and transparency, supporting whole university or institutional approaches to student success. As a result, users could turn off the adaptive settings if he/she perceives any unfair treatment.

Moreover, these systems collect data from users who may not be aware of the data being collected. Legislation such as GDPR (in Brazil, LGPD) codifies societal values in data governance and its use, reflecting a growing concern that people should control technology and its use of their data. China (2022) passed legislation for recommendation services to abide by principles of fairness, openness, and transparency, requiring explainable algorithms and prohibiting AI algorithms from offering different prices to different people based on personal data. Thus, communication between educational stakeholders and ML-based systems overcome the opacity of decision-making. In addition, their freedom is maintained with continual awareness of the amount of constraint the system is exercising over the human, and both sides can propose adjustments in response to changing conditions.

6. Discussion

In this work, we argued that intelligent educational systems could present unintended consequences by considering the reality of their developers instead of their users, which may cause biases towards different groups of people. We must remember that bias, and fairness, to a broader extent, is a social construct. As such, it should be analyzed critically:

what is unfairness? Unfair to whom? What now? As argued by Hutchinson & Mitchell (2019), the technical definitions of fairness and mitigations should not be far apart from the public's perception of fairness to obtain the political will to use scientific contributions in advance of public policy.

Critical theories are evaluated by three elements: insight, critique, and transformation (Myers & Klein, 2011). This work provided insights on how EP offers lenses to explain bias in educational ML-based applications. Next, we critiqued current development efforts challenging how upstream sources must be aware of the EP principles. Finally, as a transformation step, we presented two design propositions derived from four principles of emancipatory pedagogy related to upstream sources of bias that should be considered when developing intelligent educational systems. By leveraging a humanized perspective of users and promoting human agency, we show how theory-building, design efforts, and methods should be aware of unintended sources of algorithmic bias. Both design propositions are best achieved by co-design (or participatory design) with a varied population, as shown in Khosravi et al. (2022). The use of tools and frameworks may also help to alleviate this problem - but primarily for downstream sources of bias.

Three common subpopulations are commonly studied (Baker & Hawn, 2021): gender, race/ethnicity, and nationality. Karumbaiah (2021) points out other subgroups that are impacted by the design of these systems but are understudied: disabilities, urbanity, socioeconomic status, international students, etc.

This work presents some limitations. Providing only one theoretical kernel may leave many other aspects out. Here, we only argue for emancipatory aspects of machine learning related to algorithmic bias. Also, incorporating different steps into the ML pipeline introduces additional costs and complexity, which may deter innovation with real-world applications. However, we argue that high-stakes educational applications, particularly from public sectors, should conform to these design propositions. In addition, design propositions should be field tested through pragmatic validation, which is not contemplated in this paper. However, this theoretical paper provides support for future work on this validation. We hope this study can provide initial efforts on the study of algorithmic bias in education in Brazil - a country with a colonial past and unequal opportunities for disadvantaged groups of people - using the theory of a Brazilian educator.

References

Baker, R., Hawn, A. (2021). Algorithmic bias in education. International Journal of Artificial Intelligence in Education. DOI: 10.1007/s40593-021-00285-9.

Blikstein, P. (2018). Time to Make Hard Choices for AI in Education. Keynote talk at the 2018 International Conference on Artificial Intelligence in Education.

Boza, P., Evgeniou, T. (2021). Implementing AI Principles: Frameworks, Processes, and Tools. Working paper. Social Science Research Network.

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

China (2022). Internet Information Service Algorithmic Recommendation Management Provisions (English translation). Retrieved from: https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022.

Coleman, J. S. (1966). Equality of Educational Opportunity, Government Printing Office, Washington, DC (1966).

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.

Denyer, D., Tranfield, D., Aken, J. E. (2008). Developing design propositions through research synthesis. Organization Studies, vol. 29 (3), p. 393-413. DOI: 10.1177/0170840607088020.

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., Cassell, J. (2013). The Effects of Culturally Congruent Educational Technologies on Student Achievement. In Artificial Intelligence in Education, p. 493-502.

Francis, P., Broughan. C., Foster, C., Wilson, F. (2020). Thinking critically about learning analytics, student outcomes, and equity of attainment, Assessment & Evaluation in Higher Education, 45:6, 811-821, DOI: 10.1080/02602938.2019.1691975

Freire, P. (1970) Pedagogy for the Oppressed. Continuum International Publishing Group, New York, NY, USA.

Gardner, J., Brooks, C., Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the International Conference on Learning Analytics & Knowledge, 225-234.

Gregor, S., and Jones, D. 2007. "The Anatomy of a Design Theory," Journal of the Association for Information Systems (8:5), pp. 312-335.

Gulson, Kalervo N./Webb, P. Taylor (2017): 'Life' and education policy: intervention, augmentation, and computation. In: Discourse: Studies in the Cultural Politics of Education, 39:2, 276–291.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency (FATED), p. 49–58. DOI: 10.1145/3287560.3287600.

Kane, G. C., Young, A. G., Majchrzak, A., Ransbotham, S. (2021). Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants. MIS Quarterly, (45: 1) pp.371-396.

Kao, G; Thompson, J. . (2003). Racial and ethnic stratification in educational achievement and attainment. In Annual Review of Sociology, 29 (1), p. 417-442. DOI: 10.1146/annurev.soc.29.010202.100019.

Karumbaiah, S. (2021). The Upstream Sources of Bias in Adaptive Learning Systems.

Karumbaiah, S.; Brooks, J. (2021). How Colonial Continuities Underlie Algorithmic Injustices in Education. Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT), 2021, pp. 1-6, DOI: 10.1109/RESPECT51740.2021.9620605.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., Gasevic, D. (2022). Explainable Artificial Intelligence in education. Computers and Education: Artificial Intelligence vol. 3, 100074. DOI: 10.1016/j.caeai.2022.100074.

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

Lee, H., Kizilcec, R. F. (2020). Evaluation of Fairness Trade-offs in Predicting Student Success. EDM 2020 Fairness, Accountability, and Transparency in Educational Data (FATED) Workshop.

Kizilcec, R. F., Lee, H. (2022). Algorithmic Fairness in Education. In W. Holmes & K. Porayska-Pomsta (Eds.), Ethics in Artificial Intelligence in Education, Taylor & Francis.

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. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 444-460).

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.

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.

Myers, M. D., & Klein, H. K. (2011). A set of principles for conducting critical research in information systems. MIS Quarterly, 35(1), 17–36.

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, R., Baker, R., Mendez, G. R., Castro, M. J., Laurentino, T., Carvalho, A. (2012). Collaboration in cognitive tutor use in Latin America: field study and design recommendations. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 1381–1390. DOI: 10.1145/2207676.2208597.

Shum, S. J. B. (2018). Transitioning Education's Knowledge Infrastructure. Keynote talk at the 2018 International Conference of the Learning Sciences (ICLS 2018).

Suresh, H., Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms,

Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, --, NY, USA. ACM, New York, NY, USA 9 Pages. https://doi.org/10.1145/3465416.3483305.

Williamson, K., Johnson, G. (2018). Research methods - Information, systems, and contexts. 2nd edition. Elsevier.

Young, A.G. 2018. "Using ICT for Social Good: Cultural Identity Restoration Through Emancipatory Pedagogy," Information Systems Journal (28.2), pp. 340-358.