Ethical Dilemmas in Current Uses of AI in Science Education

Julia Svoboda Gouvea*
Department of Education, Tufts University, Medford, MA 02144

ABSTRACT

The purpose of the Current Insights feature is to highlight recent research and scholarship from outside the LSE community. In this installment, I review a series of recently published articles which examine ethical dilemmas concerning the use of artificial intelligence (AI), more specifically machine learning, in science education. The articles in this set are intended to stimulate discussions about whether and how AI can and should be used in education research.

INTRODUCTION

The recent proliferation of artificial intelligence (AI) across many different contexts has been met with controversy. While some emphasize the potential benefits of the new technology, others have raised serious concerns about its potential to perpetuate harm. These conversations remind us that research always involves ethical choices, and that as a research community it is important to have critical discussions about the impact of our work.

In this installment of Current Insights, I review recently published articles which examine ethical dilemmas concerning the use of AI, more specifically machine learning (ML), in science education. ML is a subclass of AI that involves developing algorithms that can learn from statistical associations in large datasets to perform functions like predicting text or classifying images. At the center of the controversy in science education is the specific use of ML as a tool for assessing student work.

To begin, I present an overview of an essay by Cheuk (2021) that lays out the connection between ML-based assessments and racism. This essay helps set the context for the ethical controversy surrounding a study by Zhai and colleagues (2022) that explored the use of ML to automate scoring of work by middle school science students. Last, I present a position paper by Kubsch *et al.*, (2023) who argue that the field of education research needs to move beyond uses of ML for assessment to explore more creative goals and workflows while minimizing harm.

ON POSITIONALITY

Before summarizing the papers, it is important to consider how researcher positionalities may inform participation in these conversations. Of the articles in this set, only Cheuk's essay (2021) provides an explicit positionality statement in which she describes how her own experiences as an immigrant and former English learner position her to notice examples of linguistic bias that others without such life experiences might miss.

Another aspect of positionality concerns who has the power and privilege to speak out in controversial conversations. Raising concerns about AI in education research can come with substantial professional risk, especially for emerging scholars who may occupy less powerful positions on research teams. It is also worth attending to, as Cheuk does, who stands to benefit personally, professionally, and/or financially from a rise in the popularity of AI tools in education and the creation of new subfields of education research. As a tenured professor who has been a part of a collaborative

CBE Life Sci Educ March 1, 2024 23:fe3
DOI:10.1187/cbe.23-12-0239

*Address correspondence to: Julia Svoboda Gouvea (julia.gouvea@tufts.edu).

project involving ML, I acknowledge the privilege to amplify concerns and criticisms in this forum.

A BRIEF INTRODUCTION TO RACISM AND LINGUISTIC BIAS IN ML ASSESSMENTS

Cheuk, T. (2021) 'Can AI be racist? Color-evasiveness in the application of machine learning to science assessments', *Science Education*, *105*(5), 825–836. https://doi.org/10.1002/sce.21671.

Cheuk's essay (2021) elaborates on how racial and linguistic discrimination can become amplified by using ML to assess student work. As Cheuk argues, bias has always plagued science assessment, and part of the problem is that ML methods inherit these biases.

In assessment development, success is often defined narrowly, privileging normative white discourse over linguistic expressions common among students from racially and linguistically marginalized groups. Such scoring reinforces biased interpretations of white students as more proficient.

When bias in assessment design meets ML, Cheuk argues, these issues can be made worse. Whereas humans have the capacity to use context to make sense of unexpected or unconventional uses of language, statistical algorithms cannot. This leads to the potential for further bias as ML is more likely to categorize such responses as irrelevant or incorrect. This sorting process can then set up a feedback loop wherein the linguistic forms characterized by AI as less proficient inform human interpretations of what proficiency looks like, furthering deficit narratives about already marginalized groups.

In addition to shedding light on the ways in which bias can and has entered into the use of automated ML assessments, Cheuk offers three tactics that can be used to mitigate the potential for harm. The first is a call to standardize procedures for documenting and sharing information about the student populations represented in datasets and any biases in performance outcomes across groups. Such practices can encourage accountability and transparency in ML research. Second, Cheuk calls for theoretical work that can support researchers in understanding and interrogating the relationships between science assessments and racial and linguistic biases. Third, Cheuk argues that critical intersectionality should be part of the practice of ML research. She highlights the work of critical scholars as collaborators on research teams and as organizers and educators who are raising public awareness about the potential for harm in AI research.

Cheuk closes by sharing her concern that, "In failing to take a critical stance about how the design and repercussions that this technology has on our students who are on the margins, educators risk being complicit in the work these machines do that continues to protect those in power and maintain systems of advantage for those who possess normative discourses in science" (p. 833). In this, Cheuk foreshadows the debates to come.

A DEBATE OVER USING ML FOR ASSESSMENT IN SCIENCE EDUCATION

In this section, I briefly summarize a scholarly debate about the use of ML to assess students in science that begins with a research article by Zhai and colleagues (2022) and is followed by a series of three published commentaries.

Zhai, X., He, P., & Krajcik, J. (2022) 'Applying machine learning to automatically assess scientific models', *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Li, T., Reigh, E., He, P., & Miller, E. (2023) 'Can we and should we use artificial intelligence for formative assessment in science?', *Journal of Research in Science Teaching*, *60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Zhai, X., & Nehm, R. H. (2023) 'AI and formative assessment: The train has left the station', *Journal of Research in Science Teaching*, *60*(6), 1390–1398. https://doi.org/10.1002/tea.21885

Krist, C., & Kubsch, M. (2023) 'Bias, bias everywhere: A response to Li *et al.* and Zhai and Nehm', *Journal of Research in Science Teaching*, (September), pp. 2395–2399. https://doi.org/10.1002/tea.21913

In their article, Zhai *et al.*, (2022) report on the use of ML to assess students' proficiency with "modeling" — a core scientific practice described in the Next Generation Science Standards (NGSS Lead States, 2013). In this exploratory work, the researchers use ML methods to score six items designed to assess one NGSS performance expectation[1] that asks students to create models (i.e., written explanations and drawings) related to how particle movement changes with temperature. For example, students were asked to construct a particle-level representation of what happens when red dye is dropped into water dishes of different temperatures and to write a description to go with it.

The team first generated rubrics that scored both drawings and text using three levels: *beginning*, *developing*, and *proficient* (p. 1775). Drawings were classified as proficient only if they depicted all the expected correct features. Similarly, written explanations needed to explicitly articulate the correct target ideas (e.g., at higher temperatures both water and dye particles will move faster). The rubrics were then used by human coders to create a dataset that could be used to train ML models to automatically score student work.

The researchers found relatively high levels of agreement between human-coded scores and ML-predicted scores (Kappa values between 0.64 and 0.82 for test sets), which they argue provides "robust evidence for the usability" of automated ML scoring methods (p. 1787).

However, they also discuss an important limitation of ML scoring revealed by qualitative analyses of mismatches between human and computer scores. The ML model was easily "confused" by (and thus scored as incorrect or irrelevant) unexpected features of students' drawings (e.g., the direction of arrows, the presence of a redundant label, or a drawing of a cup around the particles). The authors conclude that the algorithm's difficulties stemmed from heterogeneity in student expressions and suggest that tasks might need to be further constrained "to reduce the diversity of students' responses" (p. 1789).

In their commentary, Li *et al.*, (2023) point out that this limitation is an example of how the use of ML in science assessments contributes to the racial and linguistic biases described

[1] Performance expectations are statements that describe possible activities and outcomes that demonstrate evidence of how students integrate understandings of disciplinary core ideas, cross-cutting concepts and disciplinary practices described in the K-12c Science Education Standards (NRC, 2012).

by Cheuk. They argue that the inability of ML to account for a diversity of student ideas and expressions undermines its utility as a tool for formative assessment in science classrooms. Citing Furtak *et al.*, (2019), they argue that the purpose of formative assessment is to help teachers identify and engage with a broad "range of sense-making resources that students employ as they engage in scientific practices" (p. 1386). ML models that are biased towards recognizing normative expressions cannot adequately prepare teachers to do so.

Li and colleagues (2023) also challenge the premise presented by Zhai and coauthors (2022) that without help from AI, teachers will be reluctant to assign or engage with complex assignments. Instead, they argue that the effort teachers put into working to understand and see the value in students' ideas is central to the work of teaching and not something to be "outsourced." (p. 1387).

In their response, Zhai and Nehm (2023) echo Cheuk (2021) by arguing that "the limitations and criticisms in Li *et al.*, (2023) directed at AI for formative assessment can be applied to almost all assessments" (p. 1392). They disagree however, that using AI exacerbates these problems. They see potential for ML assessments to be used by skilled teachers and argue for future research to improve the validity of AI-based assessments, effectively framing the ethical dilemma as a problem to be solved with additional research.

Krist and Kubsch (2023) enter the conversation at this point to argue for a cautious approach that both takes seriously the "huge risks," particularly to marginalized populations of students, while also leaving room for "huge potential" for AI to make contributions to science education research. In so doing they pave the way for their own future work, discussed next.

A FRAMEWORK FOR DISCUSSING THE BENEFITS AND HARMS OF AI RESEARCH

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023) 'Distributing epistemic functions and tasks—-A framework for augmenting human analytic power with machine learning in science education research', *Journal of Research in Science Teaching, 60*(2), 423–447. https://doi.org/10.1002/tea.21803

Kubsch *et al.*, (2023) present a simple framework that can potentially aid researchers as they engage in decision-making conversations about whether or how to use AI.

The authors explicitly list many of the limitations and risks associated with the use of AI that have been raised by others, including exploitation of human labor, amplification of bias against marginalized groups, inequitable economic advantage, and environmental destruction (see e.g., Benjamin, 2019; Crawford, 2021). Given the large potential for harm, they argue that researchers need to carefully weigh whether and when the potential benefits are defensible. Specifically, they explore the contrast between uses of ML to *replicate* human efforts versus uses in which humans and computational tools are *critical collaborators*.

For example, they argue that while supervised ML models can learn to categorize responses according to predefined rubrics, this process only produces an increase in the quantity of data scored. When balanced against the potential for bias that devalues work that does not fit normative expectations, the high upfront human labor costs of training, and the limited scalability across contexts, they express "hesitancy" over this use.

As an alternative, they point to the potential for unsupervised ML to be used to deepen human insights by revealing unexpected patterns that would be challenging for humans to detect or by providing opportunities for humans to identify and address their own biases as analysts. In this type of work ML does not replicate human work but becomes part of an integrated workflow in which the output of both ML and human analyses are triangulated against each other.

IMPLICATIONS

While all research is entangled with ethical dilemmas, the articles reviewed here illustrate that the application of AI in science education warrants heightened consideration and care. These conversations are not just about what AI *can* do in a technical sense, but also about making ethical choices about how or if it *should* be used and to what ends. As individuals and as a community we have a responsibility to educate ourselves about the impacts of our research to avoid harm in its application. The articles in this set are presented here as a starting point for such conversations.

ACKNOWLEDGMENTS

I wish to thank past and present members of the Tufts STEM Education community for conversations that contributed to this feature.

REFERENCES

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Medford, MA: Polity Press.

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

Furtak, E. M., Heredia, S. C., & Morrison, D. (2019). Formative assessment in science education: Mapping a shifting terrain. In Andrade, H., Bennett, R., & Cizek, G. (Eds.), *Handbook of formative assessment in the disciplines* (1st ed., p. 29). New York, NY: Routledge.

National Research Council (NRC). (2012). *A framework for k-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press. Retrieved December 15, 2023, from www.nap.edu/catalog.php?record_id=13165