Three Paradoxes to Reconcile to Promote Safe, Fair, and Trustworthy AI in Education

Rachel Slama
RAND Corporation
Boston, MA, USA
rslama@rand.org

Amalia Christina Toutziaridi
Massachusetts Institute of Technology
Cambridge, MA, USA
amaliat@mit.edu

Justin Reich
Massachusetts Institute of Technology
Cambridge, MA, USA
jreich@mit.edu

ABSTRACT

Incorporating recordings of teacher-student conversations into the training of LLMs has the potential to improve AI tools. Although AI developers are encouraged to put "humans in the loop" of their AI safety protocols, educators do not typically drive the data collection or design and development processes underpinning new technologies. To gather insight into privacy concerns, the adequacy of safety procedures, and potential benefits of recording and aggregating data at scale to inform more intelligent tutors, we interviewed a pilot sample of teachers and administrators using a scenario-based, semi-structured interview protocol. Our preliminary findings reveal three "paradoxes" for the field to resolve to promote safe, fair, and trustworthy AI. We conclude with recommendations for education stakeholders to reconcile these paradoxes and advance the science of learning.

CCS CONCEPTS

• Security and privacy ~ Human and societal aspects of security and privacy ~ Social aspects of security and privacy • Computing methodologies ~ Machine Learning • **Human-centered computing ~ Human computer interaction (HCI)**

KEYWORDS

Education, Human-centered design, Responsible AI, Teacher Perspectives, Tutoring

ACM Reference format:

Rachel Slama, Amalia Christina Toutziaridi and Justin Reich. 2024. Three Paradoxes to Reconcile to Promote Safe, Fair, and Trustworthy AI in Education In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18-20, 2024, Atlanta, GA, USA. ACM, New York, NY, USA, 5 pages.* https://doi.org/10.1145/3657604.3664658

1 INTRODUCTION

In order for technologists to build AI models suitable for teaching and learning, they require quality data. Quality data is vital in ensuring the scalability, reliability, safety, and fairness of AI applications [18]. Audio and visual recordings of student-teacher interactions hold promise for capturing the nuanced language that expert teachers use to promote student learning, build their understanding, and correct misconceptions. AI agents trained on real student-teacher dialogue have potential to dramatically improve personalized learning.

Research dating back to the 1960s advocates for the use of recording to observe classroom behavior and provide teacher feedback [11, 16]. More recently, the Measures of Effective Teaching (MET) project sought to explore what effective teaching looks like, by collecting over 20,000 video-taped lessons [2]. The COVID-19 pandemic accelerated teacher's use of technology and normalized the recording of instruction [5, 15]. At the postsecondary level, online learning through lecture capture had already become the new norm at many universities—even after pandemic safety measures were lifted [15]. For these reasons, the benefits of recording classroom instruction may already be widely accepted by practitioners.

However, as personalized learning and online tutoring at scale open new possibilities for recording one-to-one interactions between tutors and tutees, the pedagogical benefits of recording educational interactions must be weighed alongside associated privacy risks. Classroom recordings that capture people's faces, voices, opinions, or other personal information raise questions about data retention, access, storage, and use [1].

Policy guidance in the US such as Biden's Executive Order (EO) on AI and in Europe center the role of educators in ensuring safe, fair, and trustworthy AI, but few studies have examined the feasibility of these expectations, nor educators' perceptions of these systems [6].

Exploring teacher perceptions in relation to classroom recordings is imperative, as evidence suggests that teachers play an important role in the success or failure of technology integration in educational settings [19]. Teachers tend to favor technologies that complement their existing teaching strategies and beliefs about effective education [7, 13]—further emphasizing the need to include their voices in the development of new education technology.

This work-in-progress paper is one step in the development of a stakeholder-driven framework for the ethical recording and documentation of student-teacher interactions in tutoring.

Consistent with user-centered design principles [9], we interview teachers and administrators to gather insight into

L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-0633-2/24/07.
https://doi.org/10.1145/3657604.3664658

privacy concerns, the adequacy of safety procedures, and potential benefits of recording and aggregating tutor-tutee data.

2 BACKGROUND

2.1 Large Language Models in Education

Large Language Models (LLMs), such as GPT and BERT, mark a significant advancement in the field of AI. These models are a subset of deep learning architectures that excel in generating human-like content by learning from extensive datasets. Unlike traditional machine learning which relies on pre-specified features, deep learning models autonomously determine relevant features from the data, allowing them to capture complex patterns and nuances in human language; that said, the effectiveness of such models is contingent upon the quality and breadth of the training data used. This necessity for diverse and representative datasets is especially critical when the application involves human-to-human interactions, such as those in educational settings [8].

Extensive research highlights the importance of language in educational contexts [17]. For instance, teacher conversations are laden with specialized vocabulary, nuanced expressions, and instructional techniques that are crucial for effective teaching and learning; as shown by Smith et. al. [2018], the language of teacher comments has a direct impact on student performance. By integrating data from teacher interactions into the training datasets, LLMs can learn to recognize and replicate the various ways in which information is communicated in educational contexts. Beyond linguistic features, teacher conversations embody a range of interactional dynamics–including responsive dialogue, turn-taking, and emphasizing among others. Training on these aspects helps LLMs understand the rhythm and flow of classroom communication, enabling them to participate more naturally in educational dialogues [12, 14]. This data is crucial for building AI systems that can not only assess the current state of learning but also predict and facilitate the next steps in a pedagogically sound manner. Moreover, Prieto et al. [2018] argue that pedagogically meaningful and productive teaching incidents are invaluable for AI-based educational systems. These incidents provide real-world examples from which AI can learn. For instance, understanding the cues that indicate a need for intervention or an opportunity to extend learning can be gleaned from teacher data.

In short, to create tools that genuinely enhance teaching and learning, there must be a symbiotic relationship between AI developers and educational professionals.

2.2 Educator Role in Review of AI Systems

In October 2023, President Biden issued an Executive Order on Safe, Secure, and Trustworthy Development and Use of AI [3]. One of the guiding principles is the idea that for society to harness the benefits of AI for good, stakeholders must mitigate its substantial risks. In the context of education, these risks may include exacerbating societal harms such as discrimination, bias, and disinformation. Further, the EO states that there is a collective responsibility across government, the private sector, academia, and civil society to mitigate these risks.

Accordingly, the EO tasks specific federal agencies with a set of charges and deliverables and timeline by which these tasks must be completed. By next fall (one year from the Executive Order), the Secretary of Education is to develop resources, policies, and guidance on the responsible development and deployment of education Among other tasks, education stakeholders must figure out the "appropriate human review of AI decisions, designing AI systems to enhance trust and safety and align with privacy-related laws and regulations in the educational context, and developing education-specific guardrails." The present study is one step towards understanding the degree to which key education stakeholders are currently positioned to review AI and data systems that underpin the tools they use in classrooms.

AI reflects its builders, users, and the underlying data on which it is built. In that vein, LLMs composed of teacher data have the possibility of improving existing intelligent tools, but also the risk of perpetuating existing bias and harms, particularly against marginalized students.

In 2022, the European Union Commission published a set of ethical guidelines on the use of AI and data in teaching and learning, focusing specifically on the role of educators [6]. For trustworthy AI systems, the report provides a set of guidelines related to "human agency and oversight" [6]. Noteworthy are the focus on ensuring that the guidance suggests a "teacher in the loop" while the AI system is being used and the teacher is able to notice anomalies or possible discrimination and intervene accordingly.

2.3 Research Questions

Given the critical role that educators are expected to play in data and AI systems, we sought their perspectives on the following questions:

- What concerns do you have about tutoring sessions being recorded, shared with researchers, and made publicly available? What could mitigate your concerns?
- What privacy protections would you expect?
- What data, if any, do you think should be included in these recordings?
- What are the potential benefits, if any, of recording and sharing your data with researchers? What types of research questions should researchers explore?

3 METHODS

Developing and piloting an interview protocol. We recruited a seven-member working group of educators and stakeholders to vet and pilot a scenario-based interview protocol that will be used with a larger pool of educators and stakeholders. The results described in this work-in-progress are based on interviews with this initial group, with the intention of expanding our recruitment efforts with a broader and more diverse group.

Scenario-based interview protocol. Researchers and working group members co-designed a 45-minute semi-structured Zoom interview protocol consisting of two main parts. The first part asked participants to provide brief contextual information about their school settings and experience with digital learning platforms. The second part consisted of six hypothetical scenarios related to the recording of teacher and student data and concerns related to student and teacher privacy, what protections might mitigate their concerns, and the potential benefits and research applications of collecting such data.

C*oding and analysis procedures.* We recorded and transcribed the interviews. Two researchers coded the full set of interviews using an iterative, emic or inductive approach in Atlas.ti software. We conducted broader thematic coding to identify general barriers, concerns, benefits, potential solutions, and suggested research questions. We then conducted a second pass to refine our codes and recoded all interviews to identify illustrative quotes.

4 PRELIMINARY FINDINGS

Our preliminary findings reveal three "paradoxes" for the field to reconcile in order to promote safe, fair, and trustworthy AI. First, while developers and policymakers want educators "in the loop" in AI systems, and to intervene when they notice concerns or anomalies in the system, in practice, very few systems are set up to give teachers this type of agency. Second, to understand nuanced interactions between tutors and learners, educators highlight the need for contextual information about the learning environment; yet, collecting this type of data may not be technically or ethically feasible. Third, collecting detailed demographic information on students and tutors can help detect inequitable instruction, yet many educators had concerns about the sensitivity of this information and its potential role in perpetuating harmful narratives.

Paradox 1: Developers and policymakers want "humans in the loop" to promote responsible AI, but teachers have an entirely different kind of agency: the power to choose which tools they use, but not the power to modify or intervene in them.

While most pilot participants mentioned discretionary use of a range of digital learning tools and platforms in their classroom, very few discussed opportunities to modify or intervene in the system if they perceived any ethical, safety, or quality concerns. These participants discuss the need to understand data use and access privileges, but none mention mechanisms to be actively engaged in intervening if there is inappropriate use or performance of AI-powered tools. One former teacher noted:

> "As an educator in the school, I would, and my concern, again, would be for the child and the family. As an educator, wearing that hat, I've always been distrustful of technology and where that data goes and who owns it. So I think just naturally I would have lots of questions and I would maybe even talk to families to get them to advocate for themselves or at least ask a lot of questions about what's going to happen to that data and how is it going to be used and when and if and how it would be destroyed."

This educator is referencing their own plan to engage families, but not a mechanism that is driven by the developer of the tool. One former high school math and robotics teacher described this concern which captures a range of ways that teacher recordings could be used in educational settings:

> "I would want to know what they were studying, what their research questions were. If I was being evaluated based on the recordings or if I was being coached based on those recordings, I think I would respond to it pretty differently. If they're just doing evaluation work, I think I'd be sort of lukewarm. If they're trying to support me and find ways to build tools that provide things that are aligned with my vision of good tutoring, then I'd be really into it. And if it was something that was going into systems that I didn't feel like had good oversight, I would be strongly opposed to it. So for example, if the data was being used to train an LLM that didn't have appropriate oversight, I wouldn't want my teaching to be informing the training of the language model that I didn't have confidence was going to be doing ethical tutoring."

Paradox 2: To understand interactions between tutors and students, more data matters; but technical and privacy concerns may make that challenging.

Did it work? Did it help? These are the two questions that participants cared about. Yet, the types of data typically collected differ from those that the pilot participants thought would be best suited to answer those questions. Participants focused on the impact of individual tutor-tutee engagements on student outcomes, including achievement, attendance (e.g., did they come to school more often knowing their tutor was expecting them?), disciplinary incidents, and overall well-being. In the case of math instruction, they wanted to see students' nonverbal and body language in response to the tutor, and how the students and tutors processed and shared information (e.g., access to the white board). The former district chief technology officer noted:

> "Did it help? I mean, that was always my pet peeve with any research project. I don't just want to say that they were successful in logging into the program. That doesn't tell me anything. Did it work? Did it help? Did it cause some positive impact on that student? Maybe it wasn't that they had better test scores or that they showed better growth, but if the kids came to school and they were happy to come to school on those days because of that tutor interaction, then that's a win, and I think that's important."

Paradox 3: To detect and study inequities, researchers need access to a full set of demographic data; yet many fear that sharing this information could violate privacy concerns or perpetuate existing deficit narratives.

Several participants discussed the importance of collecting detailed information, including demographic information on learners and educators, but at the same time underscored the need for certain protections such as implementing standards for minimum cell sizes to protect the confidentiality of participants and assurances that data would not be used to promote bias.

Participants also talked about the importance of using demographic data to understand whether students are being treated differently based on their race. A former math teacher noted:

> "So I just don't see how you can get away from analyzing these videos or recordings and extracting what's instructive about them without referencing the race of the tutor and the race of the student or students and that interaction. I think it's kind of constant, even if it's something micro in how a tutor may tutor. "

5 DISCUSSION AND NEXT STEPS

We propose at least three starting points for stakeholders to address the paradoxes raised in our preliminary findings. First, incentivize developers to incorporate educators "in the loop" of their AI process at the design and implementation stages. Developers could convene teacher working groups to guide realistic and meaningful ways to build educators into technology development and adoption processes that consider the ways that technology actually gets adopted in schools. Through the Executive Order, the Secretary of Education is tasked with developing resources, policies, and guidance regarding AI which can also reinforce teacher involvement. Government and philanthropy—as funders of education research—can incentivize experimentation with different ways to gather and incorporate teacher insight. Incentives should be coordinated at the district and classroom level so that teachers are compensated for their expert guidance and not expected to take on yet another initiative.

Second, technology designers can continue to build platforms that allow schools and researchers to collect multi-modal data that answers the questions that researchers, educators, and developers care about – *did it work? Did it help?* Combining engagement, achievement, and process data for tutors and tutees could lead to powerful new breakthroughs about the science of learning. To achieve such an ambitious research agenda, data providers (e.g., tutoring companies) and data custodians (e.g., school districts) will need to collaborate to determine an agreement that protects students while allowing for the analysis of critical data linking inputs and learning outcomes. Third, researchers can promote training to identify bias in instruction and support educators in proactively countering it. Providing access to large-scale data with demographic information on students and teachers can allow researchers to explore the subtle ways that they may have differing expectations or orientations towards students based on their background and inform training.

Last, all education technology stakeholders have an obligation to support teachers in preparing for the adoption of new technologies in a way that augments the more automated, rote tasks associated with teaching, and leaves the human side of teaching to the humans. As one participant expressed, teachers may share concerns about technology de-valuing their contributions:

> "I guess one concern sometimes in the back of my head is we're all going to be replaced by smart computers and robots as far as teachers and tutors. I guess that's one negative fear I guess people have about new technologies, is, 'Am I going to be replaced by some type of piece of technology and then I'll have to reinvent myself and find a new career pathway?'"

We acknowledge the limitation of a small sample size which may not fully capture the diverse perspectives and experiences of a broader population. We intend to expand the scope of the research by conducting focus groups at K-12 schools. We aim to leverage the network of teachers and stakeholder and social media platforms to interview more practitioners. Large-scale survey data of education stakeholders' perspectives on the paradoxes raised in this paper is another promising avenue to collect insight at scale. Such efforts will enable us to gather more comprehensive data and build on our initial findings.

ACKNOWLEDGMENTS

We thank our working group participants. This research was funded in part by the U.S. National Science Foundation through Research on Learning in Formal and Informal Settings (DRL) Award Number 2321499 to PI Justin Reich.

REFERENCES

[1] Andrii Vozniuk, Sten Govaerts, Lars Bollen, Sven Manske, Tobias Hecking, and Denis Gillet. 2014. Angela: Putting the teacher in control of Student Privacy in the online classroom. *2014 Information Technology Based Higher Education and Training (ITHET)* (September 2014). DOI:http://dx.doi.org/10.1109/ithet.2014.7155683

[2] Anon. 2016. Measures of effective teaching (MET) project. (September 2016). Retrieved April 15, 2024 from https://usprogram.gatesfoundation.org/news-and-insights/articles/measures-of-effective-teaching-project

[3] Anon. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. (October 2023). Retrieved April 15, 2024 from https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[4] Asif Agha. 2006. *Language and social relations* (November 2006). DOI:http://dx.doi.org/10.1017/cbo9780511618284

[5] Ayesha Ray. 2021. Teaching in Times of Crisis: Covid-19 and Classroom Pedagogy. *PS: Political Science & Politics* 54, 1 (January 2021), 172–173. DOI:http://dx.doi.org/10.1017/s1049096520001523

[6] European Commission, Directorate-General for Education, Youth, Sport and Culture, (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*, Publications Office of the European Union. https://data.europa.eu/doi/10.2766/153756

[7] Helenrose Fives and Michele Gregoire Gill. 2015. *International Handbook of Research on teacher's beliefs*, New York: Routledge.

[8] Humza Naveed et al. 2024. *A Comprehensive Overview of Large Language Models* (April 2024). DOI:https://doi.org/10.48550/arXiv.2307.06435

[9] Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson, and Åsa Cajander. 2003. Key principles for user-centred systems design. *Behaviour & Information Technology* 22, 6 (November 2003), 397–409. DOI:http://dx.doi.org/10.1080/01449290310001624329

[10] Jo Tondeur, Johan van Braak, Peggy A. Ertmer, and Anne Ottenbreit-Leftwich. 2016. Understanding the relationship between teachers' pedagogical beliefs and technology use in education: A systematic review of qualitative evidence. *Educational Technology Research and Development* 65, 3 (September 2016), 555–575. DOI:http://dx.doi.org/10.1007/s11423-016-9481-2

[11] John Withall. 1962. Trends in Observing and Recording Classroom Interaction. *Teachers' College Journal* 34, 3 (December 1962), 91.

[12] L.P. Prieto, K. Sharma, Kidzinski, M.J. Rodríguez-Triana, and P. Dillenbourg. 2018. Multimodal teaching analytics: Automated extraction of orchestration

graphs from wearable sensor data. *Journal of Computer Assisted Learning*34, 2 (January 2018), 193–203. DOI:http://dx.doi.org/10.1111/jcal.12232

[13] M. Frank Pajares. 1992. Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research* 62, 3 (September 1992), 307–332. DOI:http://dx.doi.org/10.3102/00346543062003307

[14] Rosemary Luckin and Mutlu Cukurova. 2019. Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology* 50, 6 (July 2019), 2824–2838. DOI:http://dx.doi.org/10.1111/bjet.12861

[15] Samreen Mahmood. 2020. Instructional strategies for online teaching in COVID-19 pandemic. *Human Behavior and Emerging Technologies* 3, 1 (September 2020), 199–203. DOI:http://dx.doi.org/10.1002/hbe2.218

[16] Susan E.B. Pirie. 1996. *Classroom Video-Recording: When, Why and How Does It Offer a Valuable Data Source for Qualitative Research?* (October 1996).

[17] Tamarah Smith, Rasheeda Brumskill, Angela Johnson, and Travon Zimmer. 2018. The impact of teacher language on students' mindsets and statistics performance. *Social Psychology of Education* 21, 4 (April 2018), 775–786. DOI:http://dx.doi.org/10.1007/s11218-018-9444-z

[18] U.S. Department of Education, Office of Educational Technology, *Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations*, Washington, DC, 2023.

[19] Vincent Ruhogo Abel, Jo Tondeur, and Guoyuan Sang. 2022. Teacher perceptions about ICT integration into classroom instruction. *Education Sciences* 12, 9 (September 2022), 609. DOI:http://dx.doi.org/10.3390/educsci12090609