Community for Advancing
Discovery Research in Education

Toward Ethical and Just AI in Education Research

JUNE 2024

Acknowledgments

This brief is a product of the Community for Advancing Discovery Research in Education (CADRE) Committee on AI in STEM Education and Education Research. CADRE connects STEM education researchers funded by the National Science Foundation's Discovery Research PreK-12 (DRK-12) program to share their methods, findings, results, and products with the research and development community and the greater public, to better inform and impact policy, research and education.

Steering Committee:

Christopher J. Harris,
WestEd
Eric Wiebe,
North Carolina State University
Shuchi Grover,
Looking Glass Ventures

Authors:

Tiffany Barnes,
North Carolina State University

Sarah Burriss,
Vanderbilt University

Joshua Danish,
Indiana University

Samantha Finkelstein,
Carnegie Mellon University

Megan Humburg,
Indiana University

Ally Limke,
North Carolina State University

Ole Molvig,
Vanderbilt University

Heidi Reichert,
North Carolina State University

Suggested Citation:

Barnes, T., Danish, J., Finkelstein, S., Molvig, O., Burriss, S., Humburg, M., Reichert, H., Limke, A. (2024). *Toward Ethical and Just AI in Education Research*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

The activity was funded by the National Science Foundation, Grant # 2100823. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

All images in the report were produced by the authors. Generative AI was not used to create this brief.

Table of Contents

Foreword

By Eric Wiebe, Christopher J. Harris, and Shuchi Grover

Research and development work in artificial intelligence in education (AIED) is wide ranging and rapidly growing to support all areas of science, technology, engineering, and mathematics (STEM) teaching and learning. At the risk of hyperbole, this is potentially the most fundamentally game-changing technology for education to emerge since the internet. Building from decades of work on AI and AI-based learning and teaching technologies, the recent advances in AIED are pushing us to reimagine what is possible for STEM teaching and learning. AIED research initiatives are being speedily funded, and AIED advances are quickly becoming integrated into STEM education. It is transforming how teachers teach and how students learn. It is also transforming how education developers and researchers conduct their expansive work. There is excitement about the promise of AIED as well as growing concern that the breakthroughs in AIED are impacting everyday education practice in ways that may perpetuate long-standing biases and diminish the potential for positive outcomes.

This brief is the first in a three-part series on AIED related to STEM research, teaching, and learning. The topics address ethical approaches to AI in STEM education research, AI for STEM teaching, and AI for STEM learning. This series is sponsored by the Community for Advancing Discovery Research in Education (CADRE), a National Science Foundation-funded network for STEM education researchers endeavoring to improve STEM teaching and learning through research, development, and various information-sharing and community-building mechanisms. Researchers in the CADRE network are part of a portfolio of projects funded through NSF's Discovery Research PreK–12 (DRK–12) program. The DRK–12 portfolio is wide-ranging, with a multitude of projects that focus on applied research and development to generate innovative research-informed and field-tested tools, products, and approaches that are intended to enhance STEM teaching and learning. Over the past several years, the portfolio has grown to include an increasing number of projects that leverage AIED to achieve their goals related to teaching or learning. It is expected to continue to grow. This series has been inspired by the question, *What are the essential considerations for researchers and developers who are designing, studying, and using AI in K–12 STEM?* Our hope is that the opportunities and challenges discussed in this series will generate reflection and rich discussion for the better and support the transformative use of AI to achieve positive and wide-reaching impact for all learners.

In this first brief, *Toward Ethical and Just AI in Education Research*, the authors are concerned with the ethical reasoning and decisions made in the development, study, and use of AIED technologies. Recognizing that AIED technologies reflect both the intended and unintended biases of the designers and the wider society, they advocate for the adoption of policies and practices that prioritize ethics, equity, and justice in research and development initiatives using AIED technologies in K–12 education. In an effort to provide guidance to researchers and developers, they lay the groundwork for responsible AI research and its implementation in educational settings. This foundation draws in part from the ethics rules for research with human subjects that have guided researchers for decades, but goes beyond this to frame a more all-encompassing stance rooted in justice and equity. The authors illustrate how ethical AI research can be strengthened by building from well-established ethical principles used in research and society at large. Taking into account these principles, they propose an ethical AIED framework and a set of tools that they have found to be supportive of continuous reflection, communication, and improvement toward inclusive and equitable AIED research and development. Their guidance is in the service of ensuring that the good intentions of researchers and developers will lead to positive design decisions and actions that create inclusive AIED technology products and systems. This is a valuable contribution that encourages a shift in focus to bring ethics, justice, and the values of communities of teachers, students, and families to the forefront of research and development practice.

Introduction:
The Interplay of AI and Education

Rapid advances in artificial intelligence (AI) bring unprecedented opportunities to enhance the educational experiences of teachers and learners but also bring unprecedented challenges, questions, and potential harms. In this time of rapid technological change, it is imperative that AI in education (AIED) researchers plan for and guard against the ethical risks of AI (Blackman, 2022).

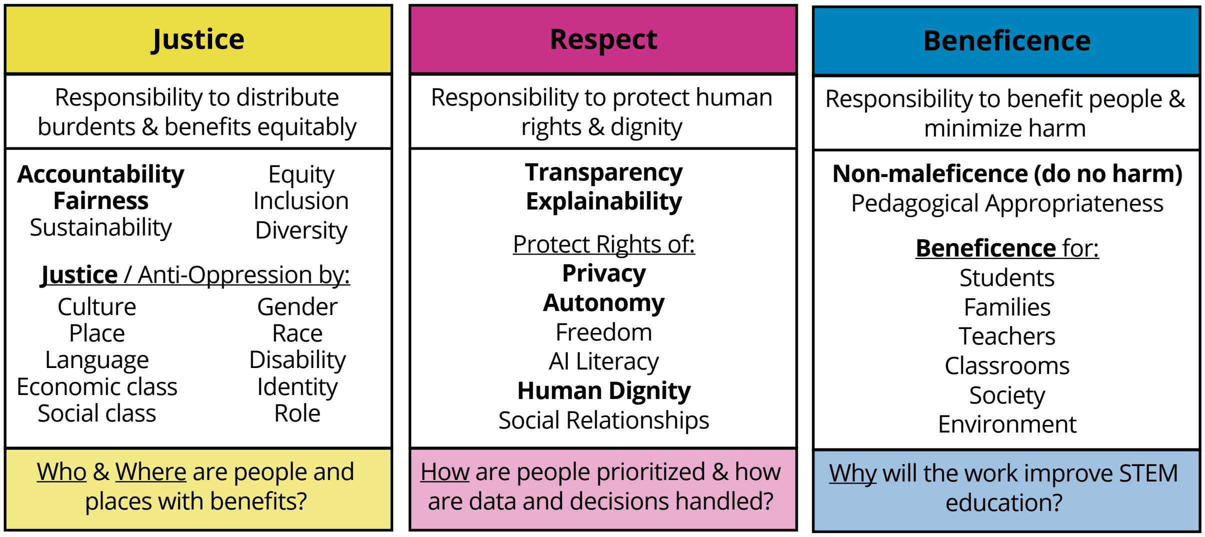
Computer science and AI researchers have significantly shaped the landscape of education. Innovative AIED technologies like intelligent tutoring systems, educational games, analytic dashboards, and new chat-based learning tools enable personalization and adaptation while automating processes like data collection, group assignment, and data analysis, facilitating work and decision-making for researchers, teachers, and administrators. In learning-sciences research, AIED technology platforms have enabled controlled studies of learning-related phenomena. Many collaborative AIED efforts among learning scientists, education researchers, and computer scientists have helped deepen the field's understanding of effective teaching and learning strategies across contexts and mechanisms. Consequently, the impact of AIED innovations extends beyond the millions of students and teachers who engage with AIED tools. AIED research and tools can also advance our comprehension of teaching and learning processes, ultimately shaping the design of future teaching and learning tools, assessment methods, teaching practices, and educational policies.

Generative AI (genAI), such as large language models (LLMs), are reshaping the landscape of education in ways that the field is still coming to understand. Teachers and students are excited by the possibilities of genAI systems because they produce relevant materials that appear to be made by intelligent humans. However, this appearance can be superficial, allowing for deep bias, mistakes, problems, and hallucinations (i.e., claims by a genAI that appear plausible but are not grounded in real data). Weidinger et al. have classified six risk areas introduced by LLMs: discrimination, exclusion, and toxicity; information hazards; misinformation harms; malicious uses; human–computer interaction harms; and automation, access, and environmental harms (Weidinger et al., 2021). However, this research is still new, and there is an urgent need for informed conversations about the limits, challenges, and dangers of AI and its new capabilities so that designers and researchers can build AIED tools that promote equitable and just educational futures.

Pursuing ethical AIED research requires collaboration among people across disciplines with interwoven technical, ethical, and pedagogical expertise. The authors of this brief are an interdisciplinary team of computer scientists, learning scientists, technology designers, and ethicists who research AI in education, equity, and justice. As AIED researchers, we benefit from doing educational research using AI and technology. We authors have worked together to construct this brief as a resource to help researchers inform the development of inclusive and equitable AIED learning experiences through an ethical AIED framework and a set of tools to support continuous reflection, communication, and improvement. In Section 1, we offer an **ethical AIED framework**, and in Section 2, we discuss three tools for guiding the design and evaluation of AIED models and their applications: **(1) reflection maps** to link theory and ethical practice, **(2) data designs** to track and visualize ethical use of data throughout the AIED software or the research project lifecycle, and **(3) model cards** for educators and researchers to evaluate, document, and communicate risks and benefits of AI models (as well as datasets and algorithms) applied in a specific educational context.

Ethical AIED Framework

Because education and AI are complex systems, ethics must be considered from multiple perspectives and at many stages of AIED work. Therefore, a framework to support ethical AIED must be much more than a checklist. Rather, it must help researchers set priorities and interrogate their work by introducing questions that can uncover potential risks, and it must be aligned with research and development processes. Figure 1 introduces our ethical AIED framework grouped by the overarching principles of justice, respect, and beneficence, adapted to guide ethical AI research (Greene et al., 2024) and for education (Roschelle et al., 2024). For each principle, the framework provides a definition (at the top), a guiding question (at the bottom), and a list of related AI ethics principles and considerations for educational contexts. Words in bold represent principles frequently included in AI ethics frameworks, including transparency, privacy, accountability, fairness, autonomy, explainability, justice, and non-maleficence (Khan et al., 2022). Additional principles were adapted from a review of state-level guidelines for the use of AI in schools (Roschelle et al., 2024), including equity, inclusion, pedagogical appropriateness, and AI literacy. The framework also highlights social, cultural, community, and societal dimensions of justice, rights, and educational roles. These considerations are complex and are best made by research teams who have read extensively from literature by leading sociologists and AI ethics researchers, such as Ruha Benjamin, Virginia Eubanks, and Safiya Noble.



Justice	Respect	Beneficence
Responsibility to distribute burdents & benefits equitably	Responsibility to protect human rights & dignity	Responsibility to benefit people & minimize harm
Accountability **Fairness** Sustainability Equity Inclusion Diversity **Justice** / Anti-Oppression by: Culture Place Language Economic class Social class Gender Race Disability Identity Role	**Transparency** **Explainability** Protect Rights of: **Privacy** **Autonomy** Freedom AI Literacy **Human Dignity** Social Relationships	**Non-maleficence (do no harm)** Pedagogical Appropriateness **Beneficence** for: Students Families Teachers Classrooms Society Environment
Who & Where are people and places with benefits?	How are people prioritized & how are data and decisions handled?	Why will the work improve STEM education?

Figure 1. Ethical AIED framework defining overarching principles and guiding questions for AI ethics in education.

This ethical AIED framework is meant as a tool for researchers to set design priorities with special considerations for educational contexts, stakeholders, and impacts. Because AI systems accomplish only what they are specifically designed to do, ethical AIED systems must include

justice as the first overarching design principle—so it is listed first. The overarching principle of **respect** centers on protecting people's rights, especially their right to learn as humans and with other humans. The overarching principle of **beneficence** includes reminders of the various potential stakeholders who may experience the benefits and burdens of research. In the rest of this section, we discuss the principles, considerations, and guiding questions within the ethical AIED framework.

Justice and AIED: Equitable Futures Are at Stake

We focus on justice as the first overarching principle for ethical AIED. Fundamentally, AI systems that are built on data, including all modern machine learning (ML) and genAI methods, are likely to perpetuate and reinforce biases present in the societal context and data from which they are produced; simply reproducing the status quo will be fundamentally unjust (Madaio, et al., 2022). Therefore, justice must be a central design focus for AIED to promote equitable and just access to education. In the ethical AIED framework, the principle of **justice** focuses on the equitable distribution of the benefits and burdens of AIED research, which in turn relates to AI ethics principles of accountability and fairness. In their systematic review of AI ethics principles, Khan et al. (2022) found **accountability** was the third-most cited principle, focusing on the accountability to safeguard justice and prevent harm and to make system decisions and take action. This responsibility is both technical and social and extends throughout the system lifecycle from design and implementation to downstream outcomes. **Fairness** was the fourth-most cited principle, focusing on the fact that AI systems are decision-making systems that can not only lead to discrimination between individuals and groups but could also foster social fairness by actively removing bias.

Within the principle of justice, we include the principles of **anti-oppression, equity, inclusion,** and **diversity**. Because students' futures depend on critical educational decisions, AIED researchers have the **responsibility** to imagine how their systems may impact these decisions, considering the guiding questions of "Who and Where" (Figure 1). That is, who and where are the people and places that benefit? This question must be considered relative to systems of power and oppression, especially considering the potential harms for people in oppressed or marginalized groups. Decision-making about student assessment, grading, and placement can have important impacts for students and their families. Producing systems that can only be accessed by well-resourced schools will exclude students and schools with disparate access to hardware, software, and the skills to use AIED systems. Equitable futures can only be achieved through intentional AIED research that prioritizes justice and anti-oppression.

Through AI built on data (e.g., ML and genAI), sources of unfairness and bias are introduced through training datasets and methods. Typically, AI systems learn labels of successful and unsuccessful behaviors and characteristics from previously collected data and situations. Bias and unfairness can be introduced through training data on the curriculum content; the characteristics, behaviors, and viewpoints of the training data populations; and the labels ascribed to them. All of these data sources may have unfair and embedded societal biases. Furthermore, most AI algorithms are trained using specific methods focused on a single objective, such as learning gains or posttest scores, which may not reflect the full educational context or potential for each student. Once these data and objectives are used to build an AI algorithm, any embedded unfairness becomes systematized and more difficult to address. There is the further risk that the conclusions that an AI algorithm derives from one population may not generalize to another population, as in typical educational research. This means that extra attention toward justice and fairness may be needed to ensure that training will lead to a supportive and understandable AI design for the targeted student population. This can be especially true for groups related to systems of power and oppression; some of these systems (e.g., income, gender, race/ethnicity, and disability status) are noted in Figure 1 under Justice.

The potential problems that may arise through the use of data and AI cannot be solved by data scientists and engineers alone; they are deeper than datasets and AI algorithms. However, the principle of **inclusion** can be applied to ensure that oppressed groups have representation within AIED datasets. Researchers must intentionally implement ethical AIED strategies and plan to mitigate bias through approaches like cross-functional interdisciplinary teams, purposefully measuring bias, and allowing for human intervention. Furthermore, many educational scholars have argued that equal material may not always be equitable and that addressing inequity might require a shift in focus. For example, many critical approaches to education note that simply including underrepresented participants is not enough; we must also help learners to recognize and combat the inequities in society (Ladson-Billings, 1995; 2014).

Respect: Transparent and Explainable AIED

The overarching AIED ethical framework principle of respect relates to the responsibility of researchers to protect people[1], their rights and their autonomy (see Figure 1). The guiding questions for this principle are “How are people prioritized?” and ”How are data and decisions handled?” Common AI ethics principles related to this question, from most common to least, are **transparency** and **explainability**, followed by **privacy, autonomy, human dignity**, and **freedom** (Khan et al., 2022). Prioritizing people within AIED research is supported by the additional principles of AI literacy—a lesser-cited principle in Khan et al.’s 2022 systematic review of AI ethics frameworks—and social relationships, a principle that centers the needs of people to learn within social relationships and contexts (Jennings & Greenberg, 2009). **AI literacy** is important for students, parents, and teachers so they can use, interpret, and make decisions about AI and AIED systems; but this is only achievable if AIED systems are designed for explainability and transparency.

Social relationships and contexts can influence learning and teacher and student well-being (Danish & Ma, 2023). AIED applications may end up playing a mediating role in these social relationships as they shape communication and make information available, and thus consideration needs to be paid to these influences. At the center of these interactions lies the teacher who often needs to work with students and their families to understand how best to support their learning. We agree with the White House Office of Science and Technology Policy (OSTP) guidelines for AI in Education, which explicitly identify the need to “always center educators” or ACE (U.S. Department of Education, 2023).

[1] In this brief, the terms “people,” “persons,” and “humans” are all used interchangeably to refer to human beings specifically and not to autonomous systems, algorithms, or AI.

AIED technologies are typically built to adapt to learners and teachers based on data and ML models, and it is important to attempt to be transparent about these adaptations by explaining how they work. However, two dimensions of AI make it difficult to achieve the principles of transparency and explainability. First, it is difficult to explain how data-driven AI models work since their algorithms are based on patterns that are not interpretable by humans. Second, AI reacts so quickly that there is little time and rarely the ability for teachers to assess or change system decisions. In high-stakes situations (e.g., decisions that bar access to challenging content or lead to disciplinary actions), it is imperative that AIED systems provide good explanations and enable human stakeholders to prevent harm. While these explanations may benefit from some level of **AI literacy**, we need to design both around the assumption that students, teachers, and parents are not AI experts and with a commitment to valuing and prioritizing positive outcomes for people.

AIED as a field is specifically interested in applying AI to learn human-centric behavioral patterns to improve learning. Respect means giving people autonomy in their decisions and control over their own data with regard to the following questions, adapted from Blackman's *AI Ethics Crash Course* (2024):

1. Transparency about data collection, usage, interpretations, and what inferences are being made about them (e.g., stereotypes or low performance)

2. Human control over how data and inferences are used (e.g., Can a teacher override a student placement into a new competence level?)

3. Assent/consent to use people's data for AI (e.g., Can an LLM use student interactions as part of its training database?)

4. Adaptation according to the amount of data shared (e.g., Do students who disallow video recording miss out on collaborative problem-solving AI suggestions?)

In making these considerations, we not only give individuals autonomy over their own data, but we also allow them the freedom to engage or disengage with AI systems during learning. In total, this principle in the framework aims to prioritize people and ensure that impactful decisions have methods for human recourse.

Beneficence: Grounding AIED Work in People's Needs

Beneficence, the responsibility to benefit people, and **non-maleficence**, to not harm, are common principles in AI ethics frameworks (Khan et al., 2022). Whereas other aspects of our framework specify the protections from harm needed by societal groups and individuals, within the overarching principle of beneficence, we focus on how AI plays a role in crafting the experience of educational stakeholders, including learners, teachers, and researchers. First and foremost, using AI instead of traditional instructional and analytic methods is a choice that **impacts sustainability and the environment**. AI models consume considerable electricity, devices, and internet bandwidth in their training and use. AI processing is increasing electricity consumption at a time when the global climate crisis is driving both a shift to carbon-free energy sources and the need for lower demand. Thus, the potential benefits of AIED must be weighed against the environmental and sustainability costs.

Furthermore, the costs and benefits should be examined with respect to **educational stakeholder roles**—students, their families, teachers, classrooms, and communities. Under beneficence, AIED researchers should consider the rights, perspectives, and voices of each educational role in the AIED design, which is a complex process. For example, considering teachers', students', and families' rights to autonomy and privacy means that they should be able to choose how and whether they use an AI-driven tool and that they have enough AI literacy to make an informed decision. Teacher, family, and student rights for social relationships may preclude using AIED that doesn't support teacher feedback, parent communications, or peer collaboration. In addition, societal and community rights may preclude using AI that differentially benefits specific groups of people based on who or where they are, intersecting with aspects of justice.

The principle of beneficence in our framework demands that AIED be grounded in both research-based theories and stakeholder perspectives on **pedagogical appropriateness**. Codesign and community-based participatory design research can make AIED systems safer, more effective, and more likely to be adopted (Centre for Social Justice et al., 2022). Students, teachers, families, and communities can provide insights into their needs, opportunities, and perceived risks. Community partnerships can achieve these same ends by connecting researchers to local organizations and resources. By application of participatory techniques, envisioned benefits in building human AI literacy, creative power, agency, and autonomy can be realized. Such techniques are important for adapting to the diverse communities in which AIED might be used. Similarly, when considering justice as a beneficence goal, Madaio, et al. (2022) suggest that researchers need new methods such as design justice and socio-technical imaginaries where marginalized communities envision liberatory learning technologies to actively counter, rather than perpetuate, structural inequity and injustice. The design justice method ensures that those who belong to multiple disadvantaged communities are heard and that created technologies are also liberatory for them (Costanza-Chock, 2020). We have purposefully tied the consideration of beneficence back to justice as AIED tools have the potential to embody and enable educational equity, or to make justice harder to achieve. We call for centering justice and learning outcomes, along with educators, in AIED research.

Ethical AIED Is a Continuous Process, Not an Outcome

The ethical AIED framework presented in Section 1 outlines the ethical principles that AIED researchers should consider when designing for ethical and learning outcomes. In this section, we present a set of tools to support AIED researchers in critically interrogating their work for ethical AIED risks and aligning their work to ethical principles as they design, implement, and study AI tools and pedagogies. **Ethical AI design reflection maps** visualize how theories about learning are embodied in AIED tool designs, become realized during anticipated processes in the learning context, and eventually lead to anticipated learning outcomes. Data designs trace how data flows through collection, interpretation, and usage for teacher- and student-facing AIED tools. **Model cards** reflect the processes used to create datasets, algorithms, and AI, and serve as researcher-facing tools to decide how and whether the model dataset should be applied (Hugging Face, 2024; Google Cloud, 2024). Each of these tools reveals the "gaps"—places where assumptions are made between phases or parts of the research, the software, the data, and the models used. These gaps are where ethical risks can be introduced into AIED research. Reflection maps help identify gaps and assumptions that connect theory to design and practice. Data designs provide complimentary tools to help identify the gaps and assumptions that connect student or teacher behavior to AIED system decisions. Model cards, in turn, help identify gaps between an externally created model and AI needs within a project. As the research, software, data, and models evolve, tools such as these should be revisited throughout the AIED lifecycle to ensure that AIED systems are designed for justice while respecting and benefiting people in alignment with the ethical AIED framework.

Design Research Using Ethical AI Design Reflection Maps

AIED research is, at its core, a human-centered design research process where human-centered user experiences, study designs, and AIED software tools are theorized to lead to learning outcomes. These study designs and tools, and the learning theories that support them, are then iteratively refined based on whether or not the proposed learning outcomes are achieved. Reflection maps help visualize the theoretical and practical connections between learning theory, tool design, use in learning contexts, and anticipated outcomes. The visualized connections between the phases of theory, design, use, and outcomes allow interdisciplinary teams of technologists and education researchers to reflect on potential gaps or assumptions that may elevate ethical AIED risks. The ethical implications at each of these phases are deeply intertwined and therefore cannot be considered in isolation. For example, while it can be tempting to treat AI technology as a "black box" within a larger educational research effort, that would naively assume that all technologies have similar ethical dimensions independent of implementation and research designs. The ethical AI framework described in Section 1 can help research teams to recognize the need to address these many dimensions of ethical

AI throughout an AIED tool's lifecycle, especially as we move from research-based AIED tool designs to realization, use in context, and AIED tool and research evaluation.

It is helpful to pair the ethical AIED framework with a model of theoretically motivated design as reified in embodied conjectures, or conjecture maps, in the learning sciences (Sandoval, 2004; 2014). Conjecture maps were developed to support design-based research (DBR), a process that aims to understand new theoretically motivated designs in rich, real-world contexts (McKenney & Reeves, 2018). Specifically, the goal of a conjecture map is to help designers articulate the overarching theoretical assumptions of the design they are aiming to study, distinguishing clearly between the underlying theory of how it will work, the instantiation of that theory into the design, and the ways that actual use in a learning context may unfold. Separating these categories makes it easier to identify how implementation results suggest success or failure in the theory, design, or implementations, or their intersections (as is often the case). While some promising efforts have been made to articulate the role of important ethical considerations within a conjecture map (Lee et al., 2022), we recognized a need to more clearly and explicitly state the key ethical decisions within a new form of ethical AI design reflection map shown in Figure 2 that integrates ideas from conjecture maps with principles within and questions inspired by the ethical AI framework.

The ethical AI design reflection map aims to provide explicit reminders of the relevant ethical considerations for designing and deploying an educational AI innovation. We have intentionally created a generic design abstraction by noting that each AIED effort will begin with considering how AI can measurably support education. To use this, designers fill out their version of this map, explicitly applying the ethics principles within our ethical AIED framework (Figure 1) by addressing the questions raised in the ethical reflection map (Figure 2) and elaborated on in Table 1. The arrows between phases highlight the direct impact of each design choice on those that follow and provide an opportunity for additional reflection. For example, decisions about how the user interface tracks student data can be made more visible to the entire AIED team, who can then discuss the tradeoffs of the tracking, how it can be made transparent and explainable to students and teachers, and how each piece of those data should be interpreted by the AIED system and researchers evaluating the data. The resulting map can be used both to guide the iterative design process and as an artifact to be shared with users or the research community to help illustrate the planned design and how it addresses ethical AIED issues. It helps structure a dialogue between stakeholders around claims made regarding educational outcomes and the inherent risks of such an approach. Each section below highlights how the different stages of the design-implementation sequence may raise ethical AIED questions.

Stage 1: Overarching Project Design

Researchers should start by articulating the overarching idea behind their design based on prior research and ask whether AI is necessary and worth the expected costs and risks. This includes considering the **beneficence** of the outcomes by defining the target population and learning outcome goals. At this point, researchers should focus on **justice via design constraints** regarding the target population asking the ethical AIED framework questions:

who is the AIED system designed to benefit, and **where** will the system be used? This overarching design is a crucial touchstone for communicating and vetting the remaining design decisions, ensuring they are aligned with the goals and assumptions of the proposed work.

Stage 2: Interfaces: Embodiments That Leverage AI in the Design

Once the learning-related outcome goals and justice-related design constraints are set, a designer decides how to embody the intelligence that will lead to those beneficial outcomes into specific software features that meet the constraints. The decisions about how to design a software feature are related to the framework principle of respect, whose guiding questions ask how the AIED system prioritizes **people**, and how data and decisions are handled. Designers should consider potential risks regarding respect through both explicit and implicit impacts on learners, teachers, and researchers. Teasing these impacts out for each stakeholder role (as listed in Figure 1 under Beneficence) can help ensure good coverage of ethical risk questions.

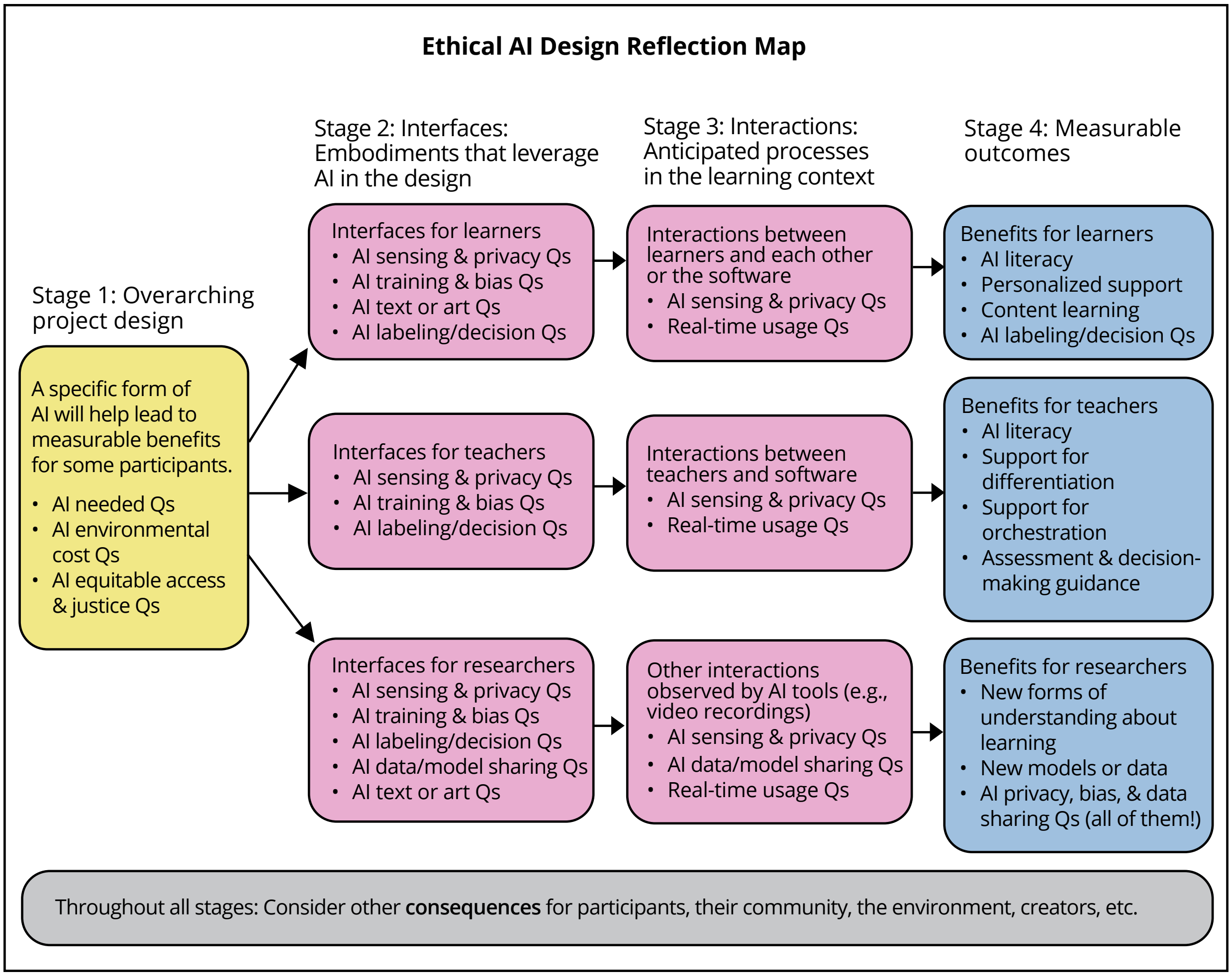

Ethical AI Design Reflection Map
Stage 1: Overarching project design
A specific form of AI will help lead to measurable benefits for some participants.
• AI needed Qs
• AI environmental cost Qs
• AI equitable access & justice Qs
Stage 2: Interfaces: Embodiments that leverage AI in the design
Interfaces for learners
• AI sensing & privacy Qs
• AI training & bias Qs
• AI text or art Qs
• AI labeling/decision Qs
Interfaces for teachers
• AI sensing & privacy Qs
• AI training & bias Qs
• AI labeling/decision Qs
Interfaces for researchers
• AI sensing & privacy Qs
• AI training & bias Qs
• AI labeling/decision Qs
• AI data/model sharing Qs
• AI text or art Qs
Stage 3: Interactions: Anticipated processes in the learning context
Interactions between learners and each other or the software
• AI sensing & privacy Qs
• Real-time usage Qs
Interactions between teachers and software
• AI sensing & privacy Qs
• Real-time usage Qs
Other interactions observed by AI tools (e.g., video recordings)
• AI sensing & privacy Qs
• AI data/model sharing Qs
• Real-time usage Qs
Stage 4: Measurable outcomes
Benefits for learners
• AI literacy
• Personalized support
• Content learning
• AI labeling/decision Qs
Benefits for teachers
• AI literacy
• Support for differentiation
• Support for orchestration
• Assessment & decision-making guidance
Benefits for researchers
• New forms of understanding about learning
• New models or data
• AI privacy, bias, & data sharing Qs (all of them!)
Throughout all stages: Consider other consequences for participants, their community, the environment, creators, etc.

Figure 2: Ethical AI design reflection map—illustrates key questions to be asked throughout the AI design and implementation process, moving from Stage 1 on the left to Stage 4 on the right.

AI needed	• What is unique about the usage of AI that could not be achieved with other technologies (or non-tech activities)? • Is this worth the costs and dangers? • Are the tradeoffs all clear to the participants?
Environmental cost	• How much computational energy does the designed AI tool or activity use? • How does this translate into electricity usage, fuel consumption, pollution, and/or water usage? • Are there versions of these tools (or design revisions that could be made) that could achieve similar outcomes with lower environmental costs?
Access	• Who will have access to what we develop? • What are the technology/bandwidth needs for users of the tools/designs? • What technological expertise do users need to be able to effectively use the tool? • How will the technology and expertise barriers limit who can use this or how? • What are the other barriers to equitable access?
AI sensing and privacy	• Have participants (and their parents, if applicable) consented to be tracked and have that data used in specified ways? • Do participants fully understand what data are being collected and how they will be used? • Is there an equitable way for participants to refuse to consent to being tracked? • Are there risks of third parties getting access to the data collected? • Are there risks of the data being used in ways that the participants did not agree to?
AI art and text	• Are these generated products (written or visual) created by an algorithm with ethically based training data? Or could the products be infused with materials and ideas whose creators did not provide consent? • Is there a way this product could have been obtained through a compensated human creator instead? • For generated text, has this product been vetted by a human reader and edited, or copied wholesale?
AI data/model sharing	• Are participants' data being used to train algorithms, and did they provide clear and active consent for their data to be used in this way? • Does the training data represent the needs and perspectives of a diverse set of users? • Has the training data been reviewed and cleaned to remove problematic, biased, and offensive content?
AI labeling and decisions	• Do those using the data (teachers, administrators, researchers) understand how to responsibly interpret the data provided by the tool? • Are the labels being produced biased towards or against certain users? • What procedures are in place to detect and remedy biased results? • Is there human input and guidance regarding how the final decision is made, and how it impacts users?
AI training and bias	• Do participants understand who has access to their data, now and in the future? • Are third parties who have access to the data committed to the same level of ethical design and data handling? • Is there an equitable way for participants to request the removal of their data from models?
Real-time usage	• Is there a procedure for handling and reporting unexpected instances of biased outputs? • Are there unexpected harms that could come to participants if the data are misused or if misguided decisions are made using their data?

Table 1: The detailed ethical questions to support the ethical AI design reflection map shown in Figure 2.

For example, consider the impact of bias in a typical AIED system for each stakeholder. First, bias in the training data may influence the content made visible to learners and the information shared with teachers, who themselves have biases in how they look at and interpret data. Second, tools that interpret these overlapping results for researchers may include yet other sources of bias. The only way to minimize the impact of bias from these different sources and their intersections is to proactively uncover and prevent them in the design process. Some of these potential opportunities and risks can be made more salient through data designs or by consulting or creating model cards (see Model Cards section, below). By challenging designers to reflect on how student-, teacher-, and researcher experiences are shaped by the AI model(s) being used, we believe it is possible to explore most of the predictable ethical consequences of the proposed design. Naturally, not all consequences are predictable, so we have included a space for "Other consequences" at the bottom of Figure 2 for both those consequences that are predicted but not easily categorized, and those that are emergent. The intention is that designers continually revise and share their map as they work but, more importantly, revisit the ethical AIED framework principles and questions as they see their software come to life and have an impact on learners.

Stage 3: Interactions: Anticipated Processes in the Learning Context

We next ask designers to reflect on how the anticipated design will impact learners in use. This may seem redundant, but the challenge is to anticipate how the AIED software interface features may interact within real learning contexts and introduce new ethical risks to justice (e.g., risks to fairness, inclusion, and equity, especially for disadvantaged[2] groups), respect (e.g., risks to privacy or lack of transparency), and beneficence (i.e., limiting teacher or student rights in pedagogically inappropriate ways). For example, we often assume small, independent, collaborative groups will use our software, but sometimes multiple groups will interact, each with a different role (Danish et al., 2020). Here, we might ask how the design, and the AI features in particular, will interact with those dynamics in new ways, and what new ethical issues and opportunities may arise. For example, in a research study on embodied learning, the software initially required physical movement. Upon reflection, the software was redesigned to allow iPad interaction for students who might be uncomfortable or unable to move in a physical space amongst their peers (Vickery, 2023), supporting justice (through fairness based on [dis] ability) and respect (through autonomy and dignity).

Stage 4: Measurable Outcomes

Finally, designers should explicitly plan how to measure or evaluate the outcomes of the design implementation, and how those measures, which are often informed by the underlying design, will impact teachers, learners, and the rest of the community. Recognizing that our measures may have positive or negative impacts on community members helps us ensure that we maintain the beneficence of the technologies in use. The simple act of measuring

[2] The terms disadvantaged and marginalized are meant to convey groups of people whose identities. or situations interact with systems of power and oppression in ways that introduce harm or reduce opportunities.

learning impacts stakeholders due to its close relationship with data collection, interpretation, and usage (see Data Designs section, below). It follows that using AI to measure and interpret learning will naturally have built-in assumptions, strengths, and weaknesses, each of which may be related to ethical principles of justice (i.e., fairness and accountability) and respect (i.e., transparency). These dimensions can be made salient for consideration using organizing tools such as data designs.

Acknowledging and exploring what the AI system assumptions are and how we can minimize the harms and maximize the benefits (i.e., respect, beneficence) are an important final step in considering how the entire system is designed (see Data Designs section, below, as one approach to this). For example, log files might be part of evaluating an AIED system in use. In the final phase of the reflection map, designers should re-examine the potential benefits and costs of log file usage for learners, teachers, and the community. The beneficence of an AIED tool can be clouded by the lure to drill down too tightly on its inner workings or its feature set (Danish et al., 2016). The challenge is to reflect on an AI-driven tool not only as a design but as a design *in use*, and a design that, once used, *is evaluated*. Such an approach has great potential to increase our attention to the impact of all of those many choices and the ways they interact. Interdisciplinary teams can use the ethical AI design reflection mapping process to interrogate their designs by inspecting the connections from design and theory to technology embodiment, use, and outcomes—similar to a conjecture map connecting theory to study design or a logic model connecting the phases of research.

Interrogate Ethical AIED Through Data Designs

In this section, we propose a second tool that can be used to interrogate our designs, asking how we can minimize the ethical risks of AIED tools from a data perspective. We suggest that AIED researchers also use the data application framework presented by Jin et al. (2021) as a guide for examining ethical AIED. Treating AIED tools as data applications involves considering ethical AIED framework principles in each phase of data collection, interpretation, and usage shown in Table 2. The goal of examining data in these settings is to address any issues that arise regarding the ethical AIED framework principles, including learner privacy and related issues such as transparency and accountability.

Stages in AIED Data Applications	Example Ethical AIED Considerations
Collection: What student data are being collected? • Monitoring clickstream data (e.g., which buttons students click) during interactive exercises • Logging question responses (e.g., correct/ incorrect answers) • Recording text, video, and audio of student interactions through the AIED tool or with classmates during the activity • Tracking time spent on tasks • Gathering demographic information (e.g., age, gender, ethnicity) • Noting seating positions in the classroom • Recording instances of hint requests	• **Justice – Fairness, equity, and cultural bias:** No data for students whose names have special characters like apostrophes or dashes • **Respect – Autonomy:** Student forgets password, teachers can't log them in • **Justice – Anti-oppression, inclusion, equity, and fairness:** student participates in an unexpected way (linguistic: non-standard English; behavioral: not using controls; cultural: not understanding directions). • **Beneficence – Technical Design Constraints:** More storage to record full answers and more difficult to analyze, but reduces ability to recover from errors
Interpretation: Where are the collected data being used as a proxy for a meaningful variable? • Interpreting text inputs to understand students' reasoning or problem-solving approaches • Analyzing behavioral cues (e.g., hesitation, frustration) during interactions with learning materials • Assessing engagement levels based on interaction patterns (e.g., rapid clicking, prolonged inactivity) • Identifying misconceptions or areas of difficulty through error patterns in responses	• **Justice – Inclusion, equity, and culture:** Failure of representative and inclusive AI training data and human perception bias leading to inequitable differences in model accuracy, including reduced accuracy for Black faces (Bacchani, 2019) and the use of non-standard English (Lawrence, 2024) • **Respect – Interpretive leaps:** Classifications can be wrong because they may be too simplistic or not combine all contextual factors (e.g., three mistakes in a row may reflect misunderstanding, or a student could be gaming the system to get a hint)
Usage: What interferences are being performed by the AIED system as a result of data interpretations? • When to provide assistance • What type of assistance to give • Adjusting problem complexity, choosing problems • Flagging student behavior as inappropriate	• **Justice – Inclusion, equity, and language:** LLM/ AI may choose shallower content based on linguistic style and/or other (mis)classifications, precluding learning of more advanced content • **Respect – Transparency:** Student surveillance (e.g., reports of "inappropriate" system behavior may be provided to the principal/ parent)

Table 2: Data design map with example ethical considerations, colored by overarching principle justice [yellow], respect [pink], and beneficence [blue].

Utilizing Data Designs to Investigate Bias

Ethical AIED must intentionally design data applications to adhere to the principles of justice, respect, and beneficence. Because education is intricately linked with existing power structures, whenever an AIED tool collects, interprets, or uses data, that tool will either reinforce or undermine those structures (Madaio, et al. 2022). As a particularly salient example, we can look at how unintended bias can be introduced into AIED systems, including:

- Linguistic Bias: The hidden curriculum of schooling enculturates students to avoid language usage that varies from the "standard" (Champion et al., 2012). There is extensive documentation that students with non-standard usage are misattributed as having lower intelligence and lower competence, and are at greater risk of committing a violent crime (Hofman et al., 2024).
- Behavioral Bias: There is well-documented evidence of white teachers incorrectly interpreting social and behavioral signals among Black students as signs of disobedience or aggressiveness, leading to biased evaluations and disproportionate discipline (Cullinan & Kauffman, 2005; Davidson et al., 2020; Douglas et al., 2008; Downey & Pribesh, 2004; Kunesh & Noltemeyer, 2019; McGrady & Reynolds, 2013).
- Cultural Bias: Assessment of student competencies (by teachers and those within AI models or AIED tools) are influenced by implicit biases based on language and behavior as above, and by explicit metrics that privilege the experiences and learning behaviors of the dominant culture. Students may engage differently with the same AIED tool (Mittelmeier, 2017) based on different cultural norms (Ogan et al., 2015) or differences in social and motivational buy-in (Finkelstein et al., 2012; Lee & Anderson, 2009).

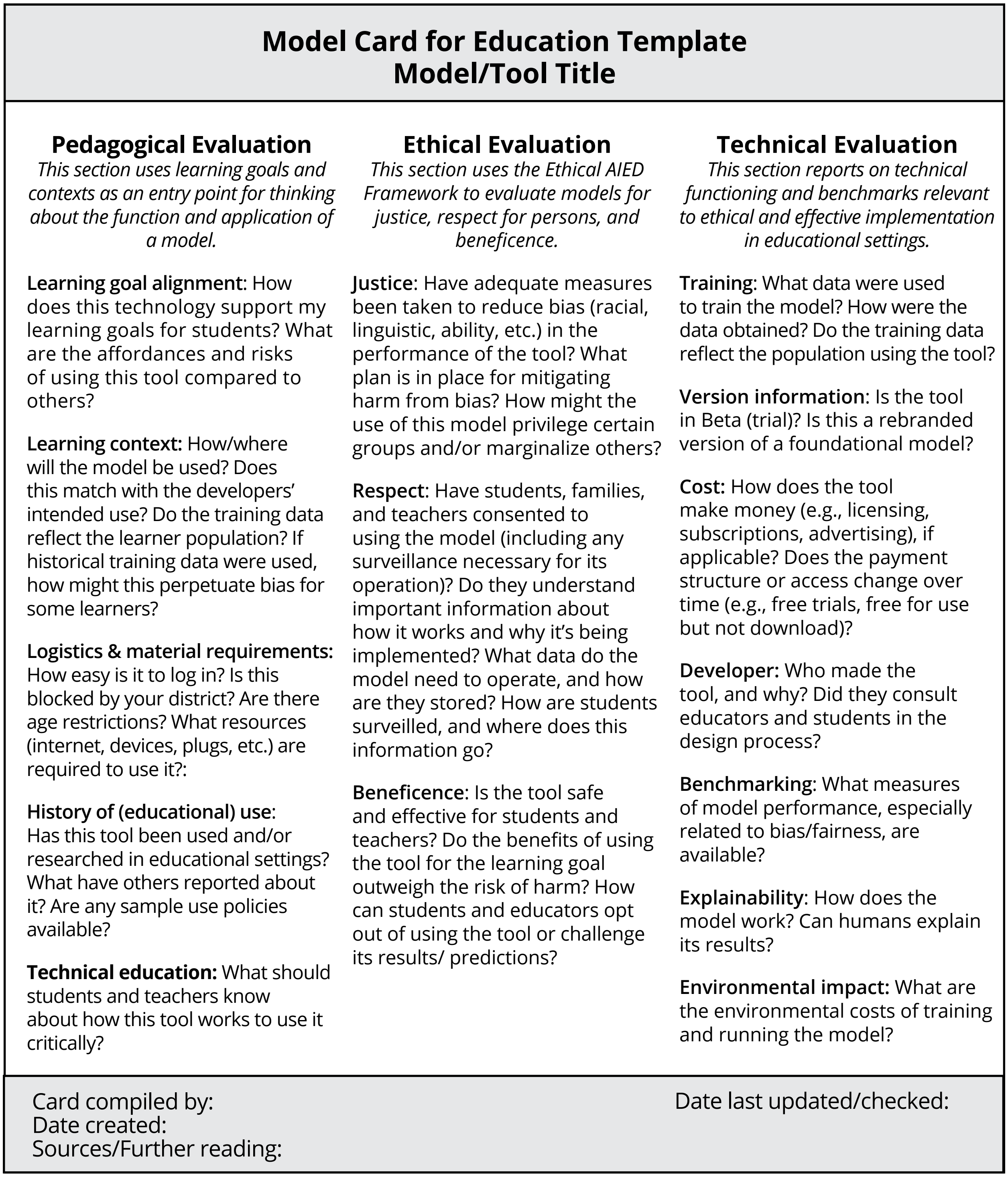
With an eye out for linguistic, behavioral, and cultural bias, we can identify the ethical risks at each stage of AIED system data collection, interpretation, and usage, and plan to avoid or mitigate these risks. This is just one example, as ethical AIED principles apply anywhere our AIED system data can be used.

Minimizing Interpretive Leaps to Respect People

AIED system design decisions that minimize interpretive leaps between data analysis and its use address respect through increased levels of **transparency, explainability**, and **autonomy** ranging from direct input to rule-based to model-based decisions. *Direct-input decisions* are based on human choices (e.g., allowing students and teachers to select problems or difficulty levels). *Rule-based decisions* are based on predefined rules (e.g., providing a hint after three wrong answers or advancing a student to the next difficulty level when they complete the current level). *Machine-learning (ML) model-based decisions* map data to decisions based on patterns in training data (e.g., a system providing a hint or increasing difficulty based on models learned from prior student work (Mostafavi & Barnes, 2017). Minimizing interpretive leaps can provide the most transparency, explainability, and autonomy but most often can be balanced with beneficence. For example, ML-based models designed to automatically predict and provide help can improve learning compared to relying on student requests (Maniktala, et al., 2022), but they are not easily explained to teachers or students. However, ML model findings do not necessarily translate from one population to another (Ocumpaugh, et al., 2014).

Model Cards

Model cards are yet another approach to structured reflection that can help researchers choose which AI models to use in their studies or to evaluate the ethical impact of the AI systems that they create (Mitchell, 2019).



Model Card for Education Template
Model/Tool Title

Pedagogical Evaluation

This section uses learning goals and contexts as an entry point for thinking about the function and application of a model.

Learning goal alignment: How does this technology support my learning goals for students? What are the affordances and risks of using this tool compared to others?

Learning context: How/where will the model be used? Does this match with the developers' intended use? Do the training data reflect the learner population? If historical training data were used, how might this perpetuate bias for some learners?

Logistics & material requirements: How easy is it to log in? Is this blocked by your district? Are there age restrictions? What resources (internet, devices, plugs, etc.) are required to use it?:

History of (educational) use: Has this tool been used and/or researched in educational settings? What have others reported about it? Are any sample use policies available?

Technical education: What should students and teachers know about how this tool works to use it critically?

Ethical Evaluation

This section uses the Ethical AIED Framework to evaluate models for justice, respect for persons, and beneficence.

Justice: Have adequate measures been taken to reduce bias (racial, linguistic, ability, etc.) in the performance of the tool? What plan is in place for mitigating harm from bias? How might the use of this model privilege certain groups and/or marginalize others?

Respect: Have students, families, and teachers consented to using the model (including any surveillance necessary for its operation)? Do they understand important information about how it works and why it's being implemented? What data do the model need to operate, and how are they stored? How are students surveilled, and where does this information go?

Beneficence: Is the tool safe and effective for students and teachers? Do the benefits of using the tool for the learning goal outweigh the risk of harm? How can students and educators opt out of using the tool or challenge its results/ predictions?

Technical Evaluation

This section reports on technical functioning and benchmarks relevant to ethical and effective implementation in educational settings.

Training: What data were used to train the model? How were the data obtained? Do the training data reflect the population using the tool?

Version information: Is the tool in Beta (trial)? Is this a rebranded version of a foundational model?

Cost: How does the tool make money (e.g., licensing, subscriptions, advertising), if applicable? Does the payment structure or access change over time (e.g., free trials, free for use but not download)?

Developer: Who made the tool, and why? Did they consult educators and students in the design process?

Benchmarking: What measures of model performance, especially related to bias/fairness, are available?

Explainability: How does the model work? Can humans explain its results?

Environmental impact: What are the environmental costs of training and running the model?

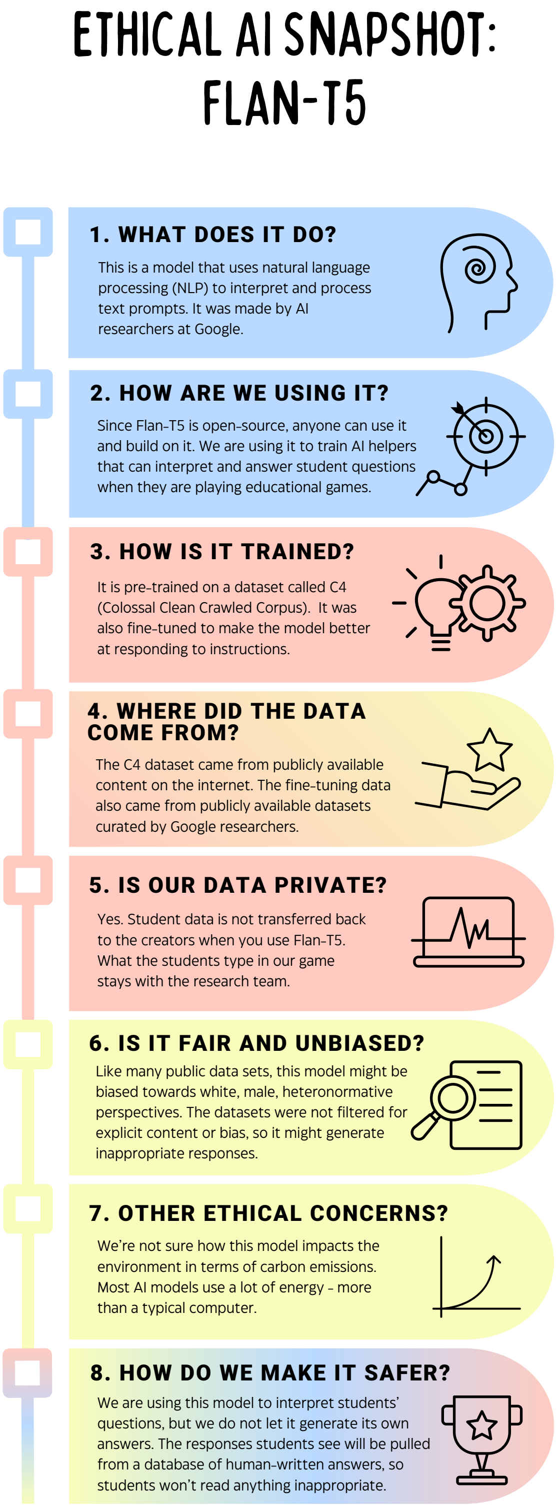
Card compiled by:
Date created:
Sources/Further reading:

Date last updated/checked:

Figure 3: Detailed model card template with guiding questions; adapted to Ethical AIED Framework.

To tailor a model card to AIED, we offer two different sample tools: a full model card template in Figure 3 and a "snapshot" model card applied to a specific AI algorithm (Flan-T5) in Figure 4. These model cards center on AI ethics, learning goals, and accessible technical explanations. In both the snapshot and template, there is a blend of technical and ethical information; for example, AI model cards for ethics should detail the model training data. They may answer key questions related to justice such as, "Was the data obtained ethically?" "Does using the data require consent or violate privacy?" "Does the data used perpetuate biases?" Additionally, if you are using a model that you do not own, it is always important to consider if data entered into the system by students and teachers is being transferred back to the creators, violating the principle of respect through privacy and consent, and model cards are one way to help do that.

The ethical AI snapshot model card (Figure 4; reproduced with permission from EngageAI Institute) is an educator-designed tool meant to guide and inform fellow educators in evaluating AI tools for use in educational settings. It is intended to help in some or all of the following ways: (1) as a guide for structuring the process of creating an education-focused model card for extant models; (2) as a thought partner or goal-setting/planning device for creating ethical AI tools for education (e.g., researchers and developers may think of how they might fill out the card as they design, test, and refine their models); and (3) when filled in, as a product that contains collected information that may be consequential for evaluating the application of an AI model for educational purposes. For example, given that the White House *Blueprint* (White House OSTP, 2023) does not specifically address the needs and concerns of students, parents, and teachers,



ETHICAL AI SNAPSHOT: FLAN-T5

1. WHAT DOES IT DO?

This is a model that uses natural language processing (NLP) to interpret and process text prompts. It was made by AI researchers at Google.

2. HOW ARE WE USING IT?

Since Flan-T5 is open-source, anyone can use it and build on it. We are using it to train AI helpers that can interpret and answer student questions when they are playing educational games.

3. HOW IS IT TRAINED?

It is pre-trained on a dataset called C4 (Colossal Clean Crawled Corpus). It was also fine-tuned to make the model better at responding to instructions.

4. WHERE DID THE DATA COME FROM?

The C4 dataset came from publicly available content on the internet. The fine-tuning data also came from publicly available datasets curated by Google researchers.

5. IS OUR DATA PRIVATE?

Yes. Student data is not transferred back to the creators when you use Flan-T5. What the students type in our game stays with the research team.

6. IS IT FAIR AND UNBIASED?

Like many public data sets, this model might be biased towards white, male, heteronormative perspectives. The datasets were not filtered for explicit content or bias, so it might generate inappropriate responses.

7. OTHER ETHICAL CONCERNS?

We're not sure how this model impacts the environment in terms of carbon emissions. Most AI models use a lot of energy - more than a typical computer.

8. HOW DO WE MAKE IT SAFER?

We are using this model to interpret students' questions, but we do not let it generate its own answers. The responses students see will be pulled from a database of human-written answers, so students won't read anything inappropriate.

Technical details if you want to learn more.

Figure 4: Example model card snapshot for Flan T5, colored to represent connections to the ethical AIED framework (where blue represents beneficence, pink respect, and yellow justice; note that 8 covers all three.

we appeal to the ethical AIED framework's concept of **beneficence** (e.g., asking "Why will the work benefit education and/or students?") and listed educational stakeholders to guide parts of the "Pedagogical Evaluation" section of the template.

Importantly, the template Model Card for Education (Figure 3) is built on an understanding of the interdisciplinarity necessary for ethical and beneficent use of AI in education, drawing on a combination of technical, ethical, and pedagogical expertise. The template Model Card is structured in three primary sections, with each of those three areas of disciplinary expertise forming the lens for gathering information on the model. These are not separate pursuits but, rather, interconnected processes. For example, when we take a technical view of the model, we cannot ignore ethical and pedagogical considerations, but we do ground that section in technical language and theory.

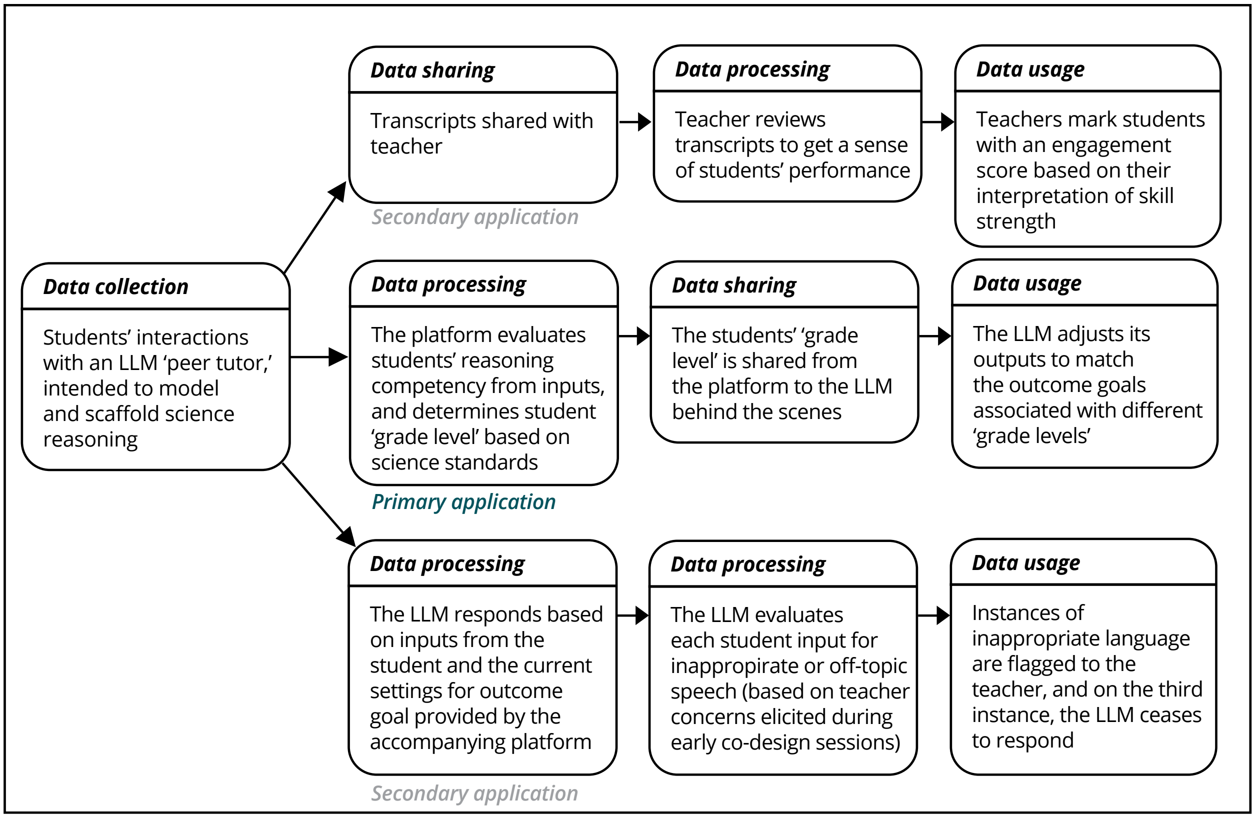
Example: Applying Ethical Tools in Context

SciBuddy is an imagined AIED tool for middle-grades science that pairs students with an LLM virtual peer collaborator. To develop SciBuddy, we would begin with an ethical AI reflection map based on Figure 2 that shows an overarching design based on social learning theories with a focus on the value of reciprocal peer tutoring (Walker et al., 2009). Answering our first question about the need for AI, we believe that AI can help each student gain the tutoring benefit in an individualized manner without over-taxing the teacher, so it seems worthwhile to explore.

In the proposed data design map shown in Figure 5, students interact with SciBuddy via a text interface similar to ChatGPT. Figure 5 shows the data map for these virtual peer interactions in the middle row. In this case, to ensure **beneficence** for learning and pedagogical appropriateness, the LLM is trained with prior on-task, successful, human peer collaboration data from students of the same age who consented to have their anonymized data used out of **respect** for their privacy. With **justice** in mind, we also want to verify that the representation, backgrounds, proportions, and behaviors of the students in the training data are similar to those who will use the tool; otherwise, the mismatch may create unintended problems such as amplifying bias.

A model card reflecting the training data features is used to reflect on these external AI and dataset aspects. The top and bottom data-map rows in Figure 5 reflect teacher codesign requests for LLM interaction transcript data, so teachers can grade student engagement (top row) and flag inappropriate language (bottom row).

The ethical AIED framework questions for Stage 2 of the ethical AI design reflection map (Figure 2) involve respect for students and teachers. Ensuring transparency and explainability involves ensuring the AI literacy rights of students and teachers to learn about how SciBuddy and LLMs work. To address this in this example, we decided to (a) start with a lesson on how SciBuddy works as an LLM to help both teachers and students, (b) make students and teachers aware of how and where chat text may be used, (c) help teachers interpret chat behaviors fairly by automated suggestions, and (d) make it easy for teachers and administrators to prevent student data from being used outside the school. This allows students to understand that they should be respectful—since teachers can read their chats—and that there could be consequences outside the system for inappropriate use. It allows teacher control over student grades and privacy.



Data collection
Students' interactions with an LLM 'peer tutor,' intended to model and scaffold science reasoning
Data sharing
Transcripts shared with teacher
Data processing
Teacher reviews transcripts to get a sense of students' performance
Data usage
Teachers mark students with an engagement score based on their interpretation of skill strength
Secondary application
Data processing
The platform evaluates students' reasoning competency from inputs, and determines student 'grade level' based on science standards
Data sharing
The students' 'grade level' is shared from the platform to the LLM behind the scenes
Data usage
The LLM adjusts its outputs to match the outcome goals associated with different 'grade levels'
Primary application
Data processing
The LLM responds based on inputs from the student and the current settings for outcome goal provided by the accompanying platform
Data processing
The LLM evaluates each student input for inappropirate or off-topic speech (based on teacher concerns elicited during early co-design sessions)
Data usage
Instances of inappropriate language are flagged to the teacher, and on the third instance, the LLM ceases to respond
Secondary application

Figure 5: SciBuddy data map with data flow stages of collection, processing/interpretation, sharing, and usage.

In using these tools, we have created a structured process, to be applied iteratively, to identify areas where ethical risks may be introduced, as well as to connect the needs of stakeholders to the needs of both the researchers and SciBuddy's design requirements.

Conclusion

Ethical practice for social justice in AIED research spans all phases of the research process, from ideation to dissemination and building on results; it is never too late or too early to start incorporating ethical principles and tools. AIED researchers must evaluate the ethics of AI tools being developed or used alongside research questions, and actively measure and mitigate how AI choices have ethically impacted teaching and learning outcomes. We recommend incorporating a structured ethics-driven reflection process throughout the AIED development and research lifecycle, using the ethical AIED framework and tools to set and achieve ethical goals. Such goals can be reached by using theory as a basis for ethical AIED research designs, interrogating our data designs for their adherence to the ethical framework principles and employing tools such as model cards to attach meaning and interpretation to datasets and models. The proposed reflection maps, data designs, and model cards for educators are tools to apply ethical AIED framework principles to guide the evaluation of ethical AI for use in classrooms and research, but they can also serve as touchstones for having conversations with key stakeholders at any point in the design, implementation, and redesign process; the AIED framework and associated tools are meant for iterative, flexible application throughout the research, teaching, and learning experience. These offerings are, importantly, not compliance measures that can be “marked complete” and forgotten but, rather, entry points into ethical practice as continuous, reflective, and recursive parts of a just and ethical AIED research process.

References and Further Reading

References

Blackman, R. (2022). *Ethical machines: Your concise guide to totally unbiased, transparent, and respectful AI*. Harvard Business Press.

Bacchini, F., & Lorusso, L. (2019). Race, again: how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society, 17*(3), 321-335.

Blackman, R. (2024). *AI Ethics: A Crash Course*. https://www.reidblackman.com/crash-course/

Centre for Social Justice and Community Action & National Coordinating Centre for Public Engagement (2022). *Community-based participatory research: A guide to ethical principles and practice* (2nd edition), CSJCA & NCCPE, Durham and Bristol.

Champion, T. B., Cobb-Roberts, D., & Bland-Stewart, L. (2012). Future educators' perceptions of African American Vernacular English (AAVE). *Online Journal of Education Research, 1*(5), 80–89.

Constanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. The MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001

Cullinan, D. & Kauffman, J. M. (2005). Do race of student and race of teacher influence ratings of emotional and behavioral problem characteristics of students with emotional disturbance? *Behavioral Disorders*, *30*(4), 393–402.

Danish, J. A., Enyedy, N., Saleh, A., & Humburg, M. (2020). Learning in embodied activity framework: A sociocultural framework for embodied cognition. *International Journal of Computer-Supported Collaborative Learning*, *15*, 49–87.

Danish, J. A., Enyedy, N., Saleh, A., & Lee, C. (2016). Designing for Activity. In V. Svihla & R. Reeve (Eds.), *Design as scholarship: Case studies from the learning sciences* (p. 26). Routledge.

Danish, J. A., & Ma, J. Y. (2023). Sociocultural and cognitive perspectives on learning: What is learning, for whom, and to what end? In R. J. Tierney, F. Rizvi, & K. Erkican (Eds.), International Encyclopedia of Education (Fourth Edition) (pp. 1–11). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.14001-1

Davidson, J., Clark, T. B., Ijames, A., Cahill, B. F., & Johnson, T. (2020). African American student perceptions of higher education barriers. *Educational Research Quarterly*, *43*(4), 59–69.

Douglas, B., Lewis, C. W., Douglas, A., Scott, M. E., & Garrison-Wade, D. (2008). The impact of white teachers on the academic achievement of black students: An exploratory qualitative analysis. *Educational Foundations, 22*, 47–62.

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, *77*(4), 267–282.

Finkelstein, S., Scherer, S., Ogan, A., Morency, L. P., & Cassell, J. (2012). *Investigating the influence of virtual peers as dialect models on students' prosodic inventory*. Multicomp Lab. http://multicomp.cs.cmu.edu/wp-content/uploads/2017/09/2012_WOCCI_finkelstein_investigating.pdf

Google Cloud. (2024). *Google Cloud model cards*. https://modelcards.withgoogle.com/face-detection

Greene, K. K., Theofanos, M. F., Watson, C., Andrews, A., & Barron, E. (2024). Avoiding past mistakes in unethical human subjects research: Moving from artificial intelligence principles to practice. *Computer*, *57*(2), 53–63.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). *Dialect prejudice predicts AI decisions about people's character, employability, and criminality*. arxiv. https://arxiv.org/pdf/2403.00742

Hugging Face. (2024). *Hugging Face model cards*. https://huggingface.co/docs/hub/en/model-cards

Jennings, P. A. & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research*, *79*(1), 491–525.

Jin, H., Shen, H., Jain, M., Kumar, S., & Hong, J. I. (2021). Lean privacy review: Collecting users' privacy concerns of data practices at a low cost. *ACM Transactions on Computer-Human Interaction (TOCHI), 28*(5), 1–55.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. (2022, June). Ethics of AI: A systematic literature review of principles and challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (pp. 383-392).

Kunesh, C. E., & Noltemeyer, A. (2019). Understanding disciplinary disproportionality: Stereotypes shape pre-service teachers' beliefs about Black boys' behavior. *Urban Education*, *54*(4), 471–498.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465. https://doi.org/10.3102/00028312032003465

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: Aka the remix. *Harvard Educational Review*, *84*(1), 74–84.

Lawrence, H. (2024). Technical and professional communicators as advocates of linguistic justice in the design of speech technologies. *Technical Communication and Social Justice, 2*(1), 1-22.

Lee, J. S., & Anderson, K. T. (2009). Negotiating linguistic and cultural identities: Theorizing and constructing opportunities and risks in education. *Review of Research in Education, 33*(1), 181–211.

Lee, U. S. A., DeLiema, D., & Gomez, K. (2022). Equity conjectures: A methodological tool for centering social change in learning and design. *Cognition and Instruction*, *40*(1), 77-99.

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2022). Beyond "fairness": Structural (in) justice lenses on AI for education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 203–239). Routledge.

Maniktala, M., Barnes, T., Chi, M., Hampton, A. J., & Hu, X. (2022). Design Recommendations for Intelligent Tutoring Systems: Volume 9-Competency-Based Scenario Design, 113.

McGrady, P. B., & Reynolds, J. R. (2013). Racial mismatch in the classroom: Beyond Black-white differences. *Sociology of Education*, *86*(1), 3–17.

McKenney, S. & Reeves, T. (2018). *Conducting educational design research*. Routledge.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Mittelmeier, J. (2017). *Supporting intercultural collaborations in blended and online settings: A randomised control trial of internationalised academic content*. British Association for International and Comparitive Education. https://baice.ac.uk/phd-abstract/supporting-intercultural-collaborations-in-blended-and-online-settings-a-randomised-control-trial-of-internationalised-academic-content/

Mostafavi, B., & Barnes, T. (2017). Evolution of an intelligent deductive logic tutor using data-driven elements. *International Journal of Artificial Intelligence in Education*, *27*, 5-36.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1974). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. DHEW Publication No. (OS) 78-0012. https://repository.library.georgetown.edu/bitstream/handle/10822/779133/ohrp_belmont_report.pdf?sequence=1&isAllowed=y

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.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, *25*, 229–248.

Roschelle, J. , Fusco, J. & Ruiz, P. (2024). Review of guidance from seven states on AI in education. Digital Promise. https://doi.org/10.51388/20.500.12265/204

Sandoval, W. (2004). Developing learning theory by refining conjectures embodied in educational designs. *Educational Psychologist*, *39*(4), 213–223.

Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, *23*(1), 18–36.

U.S. Department of Education, Office of Educational Technology, Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations, Washington, DC, 2023. Accessed online June 5, 2024 at: https://www2.ed.gov/documents/ai-report/ai-report.pdf

Vickery, M. (2023). Re-mediating technology-facilitated embodied activities at a summer camp for youth with disabilities. In Blikstein, P., Van Aalst, J., Kizito, R., & Brennan, K. (Eds.), Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023 (pp. 1394-1397). International Society of the Learning Sciences.

Walker, E., Rummel, N., & Koedinger, K. R. (2009). Integrating collaboration and intelligent tutoring data in the evaluation of a reciprocal peer tutoring environment. *Research and Practice in Technology Enhanced Learning*, *4*(03), 221–251.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Hass, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv. https://arxiv.org/pdf/2112.04359

White House Office of Science and Technology Policy. (2022). *Blueprint for an AI Bill of Rights*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Further Reading

Ethical Machines AI Ethics Crash Course
Target Audience: Those with a little time but a lot of interest in learning about the big ideas in AI ethics.
Summary: A high-level summary on the issues in AI ethics.

Artificial Intelligence Applications in K-12 Education: A Systematic Literature Review (2022)
Target Audience: Those looking for past and new ideas for research.
Summary: An IEEE systematic literature review concerning past research on AI in K-12 grades/courses, as well as recommendations for future researchers in this field.

A Holistic Approach to the Design of Artificial Intelligence (AI) Education for K-12 Schools (2021)
Target Audience: Researchers who focus on curriculum design.
Summary: The study examines challenges in designing AI-related curricula for K-12 schools, highlighting the need for sustainable approaches informed by teachers' perspectives. It proposes a holistic curriculum design model incorporating content, process, produce, and praxis, derived from thematic analysis of data from 12 schools, emphasizing six key components: AI knowledge, processes, impact, student relevance, teacher-student communication, and flexibility.

Ethical Principles for Artificial Intelligence in K-12 Education (2023)
Target Audience: Those interested in further analyses of ethical frameworks in AIED.
Summary: This paper parses AI ethics guidelines for K-12 educational contexts, finding core and newer principles in these fields referenced by other works and organizing them into a new framework for researchers.

Responsible AI and Tech Justice Guide for K-12 Education (2023)
Target Audience: Educators and students interested in critically engaging with AI and its implications in education.
Summary: This guide focuses on the core components of justice-centered computing education.

Community-based Participatory Research: A Guide to Ethical Principles and Practice (2nd edition)
Target Audience: Those interested in stakeholder input, codesign, and/or community-based participatory design research.
Summary: This guide focuses on ethical principles and guidelines for designing research by incorporating/engaging the involved communities.

Community for Advancing Discovery Research in Education

CADRE is a network for STEM education researchers funded by the National Science Foundation's Discovery Research PreK-12 (DRK-12) program. Through in-person meetings, a website, common interest groups, newsletters, and more, CADRE connects these researchers who are endeavoring to improve education in science, technology, engineering, and mathematics in, and outside of, our schools.

CADRE helps DRK-12 researchers share their methods, findings, results, and products inside the research and development community and with the greater public so that we are:

- **Better informed** about the work that is being done,
- **Continually building** on what we have collectively learned,
- **Working with our schools, communities, and policy-makers** to make our findings and products accessible and usable, and
- **Progressively able to address new and more challenging issues—**including those issues that extend beyond the limits of what any singular research project can impact.

Together, we can make a larger impact on policy, research, and education.

Contact cadre@edc.org or visit cadrek12.org for more information.

EDC.ORG