Renaissance Philanthropy

THE LEARNING AGENCY

Guidelines for the Safe and Responsible Use of AI in Ed Tech

Meg Benner, Tasha Hensley, Kumar Garg

July 2026

Introduction

As AI becomes more prevalent in schools and classrooms, developers creating AI-enabled tools and platforms must prioritize learner safety and privacy. AI offers opportunities to enhance learning in ways that weren't previously possible; yet with its many opportunities come serious risks.

A number of existing frameworks can help developers shape their approach to safeguarding; but frameworks alone are not enough. For this guidance to be actionable, ed tech developers need a resource that offers concrete, specific steps for both the design phase and the real-world application of a tool.

Drawing on our experience managing the global Tools Competition and other education R&D initiatives, The Learning Agency and Renaissance Philanthropy have identified and compiled successful safeguarding strategies from the competition's winning teams, organized them into a set of principles and corresponding actions, and included examples to make the advice more tangible.

This guide reflects proven safeguarding practices as well as emerging best practices. As AI's capabilities and risks evolve, guidelines for the safe and responsible use of AI in ed tech must be updated to reflect those changes.

Principles for Safe and Responsible AI Use

Multiple stakeholders have a role to play to ensure the effective, responsible use of AI. Policymakers must craft laws and regulations that protect students using AI-powered ed tech. Funders can support the development and scaling of responsible AI-enabled tools. Before engaging with a tool, users should understand its privacy and safety features.

Ed tech companies, with their intimate understanding of the tools they develop, have a deep responsibility to embed safety and security into the core design and application of their innovations. They must treat safe and responsible AI use as an ongoing commitment, from the creation of a tool through its deployment and maintenance, rather than as a one-time checklist to complete.

> *Ed tech companies, with their intimate understanding of the tools they develop, have a deep responsibility to embed safety and security into the core design and application of their innovations.*

From anticipating potential harms when designing the tool, to tracking and rapidly responding to problems when the tool is in use, developers must embrace a mindset of

responsible AI use and continuously take actions that prioritize the safety and security of their users. Based on our experience and research, we have identified six principles developers should follow throughout the lifespan of an AI-enabled tool:

1. Focus on educational outcomes.
2. Direct generative AI systems to support students' learning and wellbeing.
3. Ensure the privacy and security of user data.
4. Prioritize accessibility and fairness.
5. Promote transparency and explainability.
6. Give users control over their data.

This guide aims to make each principle understandable and actionable. It includes practices that are legally required in addition to other actions that are critical for protecting students, educators, and other users. This guide's checklists and illustrative examples are intended to help innovators make responsible AI use central to their product's design and implementation.

Principle #1: Focus on educational outcomes.

The chief purpose of any ed tech tool, AI-powered or otherwise, should be to improve students' cognitive and developmental wellbeing. Whether the tool focuses on helping students meet learning objectives or build durable, metacognitive, or learning skills, it must serve an educational purpose. If not, the tool risks being a distraction in the learning environment, and may ultimately compromise the quality of a student's education.
A focus on educational outcomes applies to all ed tech, whether the tool is directly aimed at enhancing instruction or indirectly supporting learning outcomes by easing an educator's administrative workload and freeing up time for more personalized instruction.

> *The chief purpose of any ed tech tool, AI-powered or otherwise, should be to improve students' cognitive and developmental wellbeing.*

Developers should:

- **Rely on research and evidence.** Tap into the existing learning science research base, as well as experts in the field, to inform the tool's theory of change, design, and implementation. Stay up to date on the latest findings as the body of research on AI and learning continues to grow. Consult resources like the U.S. Department of Education's What Works Clearinghouse, Stanford's SCALE Initiative Research Study Repository, and the Education Endowment Foundation's Teaching and Learning

Toolkit to identify and build on proven research-backed approaches. Use high-quality, evidence-based instructional materials as inputs for tools incorporating generative AI.

- **Co-design with users.** AI-enabled tools should align with the educational goals and needs of their users. That's why it's important to engage authentically with prospective users when designing the tool. Students, educators, school administrators, and families are closest to the challenges ed tech tools are intended to solve, and therefore best positioned to co-create innovative solutions. To facilitate co-design, include practitioners or students on the design team, host focus groups or design sessions with prospective users, and offer prototype demos to spark discussion and ideation.

- **Stay engaged with users.** Invite students, teachers, and other users to provide feedback throughout the tool's implementation. Offer multiple channels for feedback, such as a button for anonymous, aggregated real-time feedback (e.g., "like" or "dislike" button) and an easy-to-find form for more detailed comments. Monitor and address user feedback on a consistent basis.

- **Ensure that the tool complements and enhances, rather than replaces or conflicts with, teacher instruction.** When applied responsibly and in alignment with a teacher's instructional goals, AI can enhance student engagement and learning while easing administrative burdens. But it cannot replace a human teacher leading the classroom, as education is an inherently social process that relies on a teacher's creativity, intuition, and ability to connect with their students.

- **For student-facing tools, build in pedagogical guardrails to mitigate the risks of AI-enabled cognitive offloading.** A growing body of causal studies suggests that AI tools improve student performance in math, programming, and writing tasks while they have access to the technology; but the results are mixed when students are assessed without having access to AI support. AI tools can enrich a student's learning experience, but if not designed and implemented carefully, they can endanger deeper thinking and long-term learning gains. According to research by Stanford's SCALE Initiative, AI tools with "pedagogical guardrails" show more promising results than general purpose AI tools. These guardrails keep students engaged in learning. Tutoring chatbots with pedagogical guardrails, for instance, would use Socratic scaffolding, provide hints, or offer step-by-step reasoning rather than directly providing answers to the students. It's also important for AI tools to target students' Zone of Proximal Development — offering support that stretches them cognitively without venturing into content they cannot yet grasp.

- **Center humans in implementation.** To prevent harmful or misleading AI-generated outputs from reaching students, educators must be involved at key points in the AI workflow — an approach called "human-in-the-loop" (HITL). In the HITL model, humans bring their oversight, expertise, and final judgment to a tool's AI-generated outputs. In practice, this means educators should take on the role of AI monitor, editor, and validator. For instance, an AI-enabled tool can provide feedback on student essays but the teacher must review and edit that feedback for accuracy before students receive it. HITL should be a natural part of the workflow; it should be easy for teachers to include their own feedback in the outputs that reach students. Additionally, the tool should include mechanisms that invite educator feedback in an effort to refine its effectiveness.

- **Continuously measure and work to improve the accuracy rate of the tool's core AI task.** The core task reflects the role AI is intended to play in the tool, whether it's grading student essays, identifying reading errors in passages students read aloud, or surfacing students' math misconceptions. Define the accuracy rate and explain the measurement methodology, including the composition of the evaluation set. Set an accuracy rate goal and articulate a path to achieve it. Track the accuracy of the core AI task and report accuracy rates disaggregated by relevant subgroups (e.g., English Learners, students with Individualized Education Plans (IEPs)) to detect and address any biases. Be prepared to share these results with prospective school partners before a contract is signed.

- **Commit to evaluation and continuous improvement.** Regularly evaluate the tool's educational impacts using methods that range from rapid A/B tests to large scale Randomized Controlled Trials. Various forms of rapid testing are particularly important as AI's impact on education quickly evolves. Work with external research partners to ensure independent reviews. Use established benchmark datasets to measure model performance with standardized industry-specific data. Make any necessary adjustments, always with the goal of improving teaching and learning. Publish research results, which may include null or negative findings, to ensure users have the latest information on the tool's effectiveness.

What does this look like in practice?

MentorPRO is a research-backed mentoring platform that allows education, workforce, and nonprofit organizations to deliver high-impact mentoring at scale while maintaining a human-centered experience. Integrated within the platform is MentorAI, a 2025 Tools Competition winner, that supports mentors in real time by offering tailored conversation prompts, resources, and insights based on mentee activities and program goals. MentorAI incorporates a "human-at-the-helm" review process in which trained mentors review and either validate, adapt, or ignore AI-generated recommendations before they are delivered to students, combining automated checks with human oversight.

Principle #2: Direct generative AI systems to support students' learning and wellbeing.

Generative AI in ed tech poses unique risks for students and other users. Its outputs have the potential to be age-inappropriate, off-topic, biased, and inaccurate. When designing an educational tool and its underlying AI system, developers must mitigate each of these risks.

This means using high-quality training data and making design choices to ensure the Large Language Model's (LLM) outputs are: 1) age-appropriate (i.e., not explicit, at the right level cognitively and developmentally), 2) topic-focused (i.e., limited to the intended scope of learning), 3) unbiased (i.e., only personalizing content based on carefully selected factors like evidence of student learning or interests rather than demographic factors), and 4) factual (i.e., anchored in a knowledge base). Developers must also monitor the tool's AI outputs constantly to ensure the model is performing as intended, and address any issues that arise.

There is also the risk of a user's input being inappropriate and disruptive or harmful to the learning experience. Developers have a responsibility to plan for these types of incidents and respond swiftly when they occur.

Developers should:

- **Train and routinely test the AI model to prevent its outputs from causing harm.** Keeping generative AI conversations safe starts with training the AI model with high-quality data. Additionally, a moderation layer, which filters or flags harmful content before it's delivered, can be inserted between the LLM output and students. The model's outputs should be tested routinely, before and throughout implementation, to ensure they are appropriate, accurate, and free of bias.

- **Engage in red teaming.** This is an approach that surfaces vulnerabilities by adversarially prompting the model to act inappropriately. Fix any problems that are detected in the red teaming process. Red team routinely, in the design and implementation phases, to prevent harms before they impact users.

- **Put a content moderation plan in place to detect and respond to harmful AI outputs, as well as inappropriate user inputs.** This plan should have clear roles and responsibilities for team members who will detect and respond to incidents. It should include procedures and specific timeframes for corrective actions, as well as for user and parent/guardian notifications. Content that should be flagged in the moderation process includes profanity, threats, sexual or mature content, references to self-harm, bullying, harassment, and hate speech.

- **Use Retrieval-Augmented Generation (RAG) and thoughtful prompt engineering to minimize hallucinations and biases.** The RAG approach connects LLMs to external knowledge bases. In the case of ed tech tools, this technique links the AI model to high-quality, open-source educational curricula (e.g., OpenStax) or school districts' vetted instructional materials. By leveraging a high-quality knowledge base, the LLM is less likely to generate inaccuracies or nonsensical outputs, or perpetuate biases against marginalized groups. Through prompt engineering, set parameters for the AI model that will reduce the likelihood of hallucinations. For example, program the AI model to ground its responses in verifiable sources and to direct students to their teacher if it does not know the answer. Prompts can also be iterated on to refine model alignment.

- **Invite users to flag issues**. Build feedback loops into the tool to allow users to notify the developer about any problematic features, such as a chatbot using inappropriate or biased language. Monitor and respond to user feedback on a regular basis to ensure the safety of the tool.

What does this look like in practice?

Rising Academies (a 2021 Tools Competition winner) has created Rori, an AI-powered math tutor used in over 90 schools across Africa. Rori was designed to be used via Whatsapp on affordable smartphones so that it can reach students without access to high-speed internet or high-powered devices. Its curriculum spans grades 1 through 9 and is based on UNESCO's Global Proficiency Framework for Mathematics.

Rising Academies takes several steps to anticipate and mitigate safety risks for any student using Rori, including:

- Ensuring sessions with Rori always take place under adult supervision, with phones stored securely between sessions.
- Developing safety protocols to protect students from inappropriate messages from unknown contacts, as phone numbers may be recycled from previous users.
- Putting a robust moderation process in place for when students engage in LLM-driven conversations:
  - Student inputs are first screened for profanity through keyword matching, triggering immediate pre-written responses when detected.
  - Clean inputs then pass to AI moderation to assess if it's highly sensitive. If it is, the LLM conversation is stopped and the student receives a pre-written response and the Rising team receives an email alert so that a human can review the interaction.
  - The same process is in place to moderate AI model outputs.

Principle #3: Ensure the privacy and security of user data.

Students, teachers, and other stakeholders should be able to use ed tech without fear of their data being misused. It is essential for developers to protect their users' Personally Identifiable Information (PII), information that can directly or indirectly identify an individual. While user data is useful for personalizing tools, making continuous updates, and informing research, PII must be collected, stored, and protected with extreme care. Through

intentional data governance, developers can comply with privacy laws and, crucially, respect the dignity of the student or educator using their tool.

It's critical that developers establish and carry out policies and practices that protect user data. Their approach to data governance should not only comply with laws and regulations but keep pace with new developments in AI, even if policymaking lags. In a regularly updated data management plan, developers should articulate what constitutes student data, and include guidelines for how it will be collected, stored, used and shared; which parties have access to the data; and how long the data will be retained before deletion. Below are specific aspects of data protection that should be addressed in an organization's policies and procedures.

> *Through intentional data governance, developers can comply with privacy laws and, crucially, respect the dignity of the student or educator using their tool.*

Developers should:

- **Understand and adhere to privacy laws and regulations.** Undergo regular third-party audits to help verify the tool's compliance with privacy standards. Key privacy laws in the U.S. include the Family Educational Rights and Privacy Act (FERPA) and the Children's Online Privacy Protection Act (COPPA). FERPA governs the privacy of and access to student education records, whereas COPPA regulates the collection of personal data from children under the age of 13 by commercial websites and apps. The General Data Protection Regulation (GDPR) is the European Union's data protection law, and it is known for setting a high standard for data privacy and security. Be sure to check and follow state laws, too, as some expand on federal privacy protections and may require additional consent or contract terms. Stay attuned to the latest policy developments and make adjustments to the product as needed.

- **Collect data purposefully.** Gather user data needed for a tool's functionality, personalization, continuous improvement, and, in some cases, for approved education research purposes. Do not collect data that does not serve an educational purpose, or for the purpose of unethically selling the data for institutional or personal gain.

- **Store data securely.** Put in place strict access controls, like multi-factor authentication and secure, password-protected storage systems, to ensure only authorized individuals can access the data. Encrypt data at rest and in transit so that

unauthorized parties cannot obtain the PII. In cases of sensitive data such as socioeconomic status or Health Insurance Portability and Accountability Act (HIPAA) data, consider specific handling requirements and enhanced safety protocols. Periodically review stored data to assess what can be deleted, while also following established data retention protocols.

- **Develop and execute a robust incident response plan for foreseeable threats to data security**, such as a cybersecurity attack**.** Like the content moderation plan, it should include team members' roles and responsibilities, as well as procedures and specific timeframes for corrective actions and stakeholder notifications.

- **Share data responsibly.** When sharing data with an authorized third party or research collaborator, anonymize the data by removing PII whenever possible. Through a data use or data sharing agreement, articulate how the data can be used and specify that the third party or research partner must not attempt to re-identify the data.

- **Remove PII from any data used in a generative AI model.** PII that users share with generative AI should remain private to any session. This is an important practice for protecting the privacy of user data. Training a model with data that contains PII could lead to "PII leakage," in which the model memorizes and inadvertently shares sensitive information in its outputs. Furthermore, in some jurisdictions like the European Union (which operates under the GDPR), individuals have "the right to be forgotten," which is the right to request that their data be deleted. If their PII is incorporated into an LLM's neural network, it becomes challenging to "forget" that information without retraining the entire model. If using a third-party foundation model API, be sure to put a data processing agreement into place to protect users' PII.

- **Engage in anonymization.** This is a data privacy technique that proactively detects and anonymizes PII based on the meaning of the data, not just data that is obviously sensitive, such as student names. In this approach, personal details that learners or other users share with an AI native tool through their own writing or speech (e.g., information about a learner's home or family shared during a live chat) get replaced with contextually appropriate surrogates. Semantic anonymization preserves the usefulness

of the data for analysis while minimizing the risk of re-identification. Eedi, for example, uses the "PIIvot" approach to anonymize user data, detailed here.

- **Undergo data management audits.** Developers should have an independent third party validate their data management practices. One example is an SOC 2 compliance audit, which assesses an organization's controls related to security, availability, processing integrity, confidentiality, and privacy.

- **Offer educators data privacy support and training.** While it's incumbent on developers to protect user data, school staff also have a part to play. Through training and user-friendly resources, developers can help teachers, tutors, and other personnel use the tool properly to maintain the privacy of student data. For instance, DuoLingo offers teachers a set of resources covering topics such as student data management and privacy settings.

What does this look like in practice?

Khan Academy is a nonprofit organization that created Khanmigo, a generative AI-enabled tutor. It protects user data through a set of clear, publicly shared policies and practices:

- Data privacy. Khan Academy articulates its commitment to never selling users' personal information, its adherence to student data privacy laws, and the extra precautions it takes for learners under the age of 13 (e.g., blocking them from features that allow personal information to be shared). Additionally, Khan Academy prohibits foundation model providers from using Khanmigo user data to train their models.
- Security. Khan Academy explains their data encryption measures, routine security compliance assessments, annual external penetration test and SOC2 Type2 audits, and incident response plan with processes for initial detection and reporting, communication to impacted parties, isolation, resolution, and post-mortem lessons learned.
- Data deletion. The organization conveys the purpose of the data collected (i.e., Khanmigo conversations, user insights, and documents co-authored by the user and Khanmigo) and the period after which each type is deleted. It also explains how users can request the deletion of their Khanmigo conversations or accounts.

Principle #4: Prioritize accessibility and fairness.

AI in education must be designed and implemented with accessibility and fairness in mind to avoid reinforcing biases and exacerbating disparities. The use of AI brings unique ethical considerations and risks that developers should constantly evaluate, especially as technology advances.

Developers should:

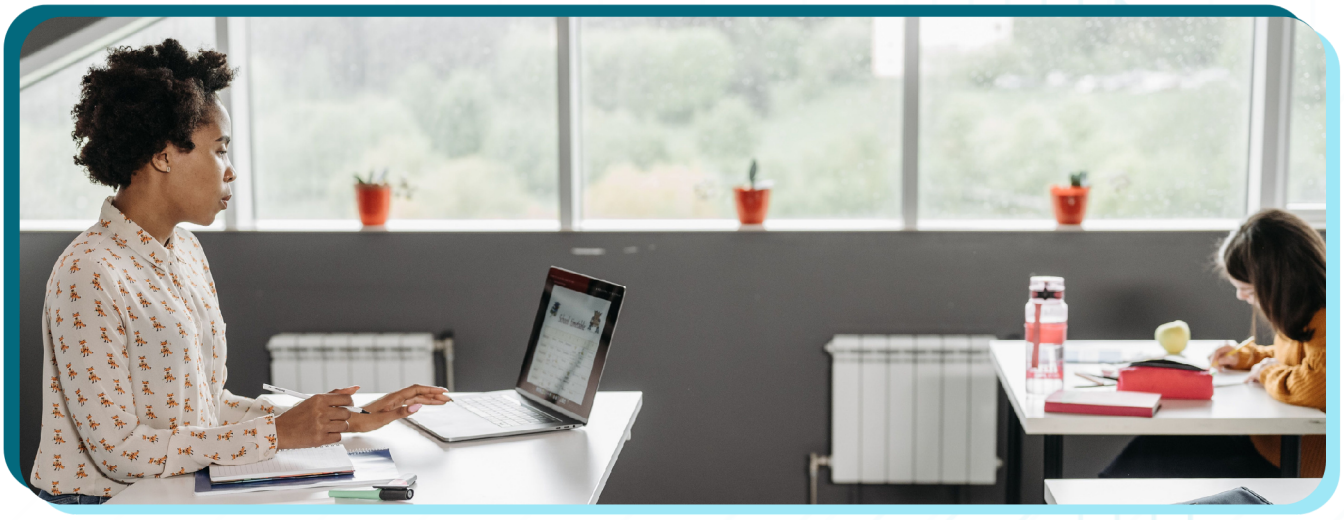
- **Consider and design for the diversity of their users.** No matter how wide or narrow the tool's target demographic, there will be diversity among the population it's intended to serve. There may be differences across age, gender, race/ethnicity, language, learning abilities, socioeconomic background, and other characteristics. There may also be variation in users' access to high-speed internet. Be aware of and design for these differences so that intended users are able to benefit from the tool, regardless of their background or abilities. If training the AI model rather than using a foundation model, ensure the training data represents the target demographic. Throughout the design and implementation processes, consider learner variability and make intentional choices to meet different users' needs (e.g., offering multiple languages, ensuring culturally appropriate content, including accessibility features). This is another reason why co-designing with users is critical; they help ensure the tool meets the diverse needs of the people it's intended to serve.

- **Monitor for bias after initial design by tracking outputs.** Regularly monitor the tool's AI-generated outputs and outcomes across any variations in contextual or demographic metadata collected by the tool (e.g., gender, race/ethnicity, English Language Learner status, IEP status, or socioeconomic status) to identify and reduce biases. Where disparities surface, revise the tool's design, workflows, or model behavior to address them. Read here about how Carnegie Learning conducted a fairness evaluation.

- **Promote accessibility for all students.** Individuals with disabilities may face barriers when trying to engage with AI-powered ed tech. As such, tools must meet accessibility standards, such as Web Content Accessibility Guidelines (WCAG) and other obligations under Title II of the American Disabilities Act. For instance, images with meaningful information should include alternative text for screen readers; color contrasts must be sufficient; and captions and/or transcripts should be available for audio. Design frameworks such as CAST's Universal Design for Learning Guidelines can help ensure that a tool's content and methods are accessible, inclusive, and appropriately challenging.

- **Test for compatibility with assistive technologies.** In the design phase, test the tool to make sure it can be used with devices or software tailored to increase functionality for individuals with disabilities (e.g., screen readers, speech-generating devices).

- **Undergo accessibility audits.** Get an independent third party, such as Digital Promise, to evaluate the tool for its compliance with accessibility standards. Create a public and easy-to-locate accessibility statement to make the results of those audits available to users and any school districts using the tool.



What does this look like in practice?

LessonLoop, a 2024 Tools Competition winner, is a professional learning platform aimed at boosting student engagement. Using AI, LessonLoop helps teachers identify engagement challenges and apply evidence-based strategies to overcome them.

In its development process, the LessonLoop team realized that, to make the classroom experience meaningful and engaging, the platform had to reflect the diverse, lived experiences of students. This led the team to create a new feature, intended to identify cultural, language, and accessibility bias in lesson content. It embeds bias detection directly into instructional workflows and allows teachers to notice and respond to issues in real time.

LessonLoop's experience developing and applying the new feature involved:

- Assembling a diverse team to examine the platform's AI-generated outputs
- Defining what "bias" means in the context of a classroom lesson
- Building a high-level prompt that screens lessons for cultural references, tone, and accessibility
- Developing and applying a rubric to evaluate lessons for bias
- Identifying and addressing patterns of bias across lessons

You can learn more about LessonLoop's efforts to detect and mitigate bias in classroom lessons on the Tools Competition blog.

Principle #5: Promote transparency and explainability.

Users deserve clear, accessible, and up-to-date information about how AI tools function. In plain language tailored for its audience (e.g., educators, students, families), developers should be transparent about the tool's intended impact; the evidence and rationale behind its design; and its risk mitigation, data privacy, and security practices. Developers should not only make sure users know when they are engaging with an AI-enabled feature, but also provide insight into the tool's AI model — which model is being used, what data it's trained on, how it works, and what risks users assume when engaging with it.

Information Schools Have a Right to Request from Vendors

Before entering into a contract with an ed tech vendor and using a specific AI-powered tool, schools have the right to request information about that tool. To promote transparency, proactively provide the following details in clear, plain language to a prospective school partner.

Information about the tool's underlying AI, including:

- Which tool features use AI
- Information about the data on which the AI model is trained
- How generative AI outputs are generated
- Any known AI model limitations
- Any measures being taken to mitigate bias in AI-generated outputs
- What human-in-the-loop practices the tool relies on

Effectiveness data, including:

- The tool's theory of change and evidence that supports its design, features, and desired outcomes
- How the tool's effectiveness is being measured
- The accuracy rates of the tool's core AI task, disaggregated by relevant student subgroups
- Results of evaluations of the tool, which may include null or negative findings
- Results of accessibility audits

Data privacy policies and procedures, including:

- What data is collected
- How data is kept secure and private
- How data is used, with whom it is shared, and why
- How long the data is retained before deletion
- How a user (or their parent/guardian) can contact the developer to request the removal and deletion of their data
- Any third party vendors involved and their data privacy policies
- Incident response protocols for threats (e.g., cybersecurity attack)
- Results of data management audits

Developers should:

- **Articulate the tool's theory of change.** Clearly convey the tool's approach and intended impact. Provide evidence that supports the tool's design, features, and desired outcomes. Explain how effectiveness is being measured.

- **Communicate policies and practices.** Publish information about the product's policies related to data privacy and any other safety measures online, and keep this information up-to-date. Ensure that the information provided is easy to find (e.g., accessible from the homepage), non-technical, and understandable to a wide variety of stakeholders, from students to policymakers. This should include information such as what data is collected and used; where the data is stored and whether models are hosted locally or deployed elsewhere; what third-party vendors are involved and what their privacy practices entail; and what choices users have in how they interact with the tool.

- **Disclose any tool features that use AI.** Make sure that users know when they are engaging with AI. If the tool has a chatbot, inform the user that they are interacting with AI, not a human. If the tool automates grading, tell the students being graded so that they can follow up with their teacher to seek clarification about or contest any grades that seem incorrect. In clear, easy-to-read language (i.e., not in fine print), let users know that AI-enhanced tools can generate content, feedback, or answers that are inaccurate, and explain what processes are in place (e.g., human review, citations, confidence indicators) to mitigate inaccuracies.

- **Strive to make the tool's underlying AI system understandable to users.** This includes explaining how recommendations are generated, providing visibility into system logic, and being transparent about the data used to train the AI model, as well as any known model limitations and error rates. Be clear about what is unknown about how the model behaves and any recommended guardrails educators can put in place.

- **Make security and privacy documentation available.** Inform users that detailed materials such as IRB data-sharing agreements and cybersecurity audits are available by request.

- **Keep users informed.** Regularly notify users of any updates to data practices or other policies, as well any new tool features. For new features, communicate clearly about what is different from standard use, why the feature is being added, and what it means for users.

What does this look like in practice?

ASSISTments (a 2024 Tools Competition winner) is a nonprofit that develops "teacher-paced, evidence-based online technology" to improve student learning. They are creating several AI-powered innovations, including a conversational AI agent that interacts with students, AI that scores and gives feedback on open-ended responses, and a data summary agent for teachers.

To help students, educators, and families understand the measures they take to deploy AI responsibly, ASSISTments added a "Trust Center" to their website, displaying compliance credentials and describing in plain language their approach to keeping users safe, including data management and privacy policies; incident response and threat detection measures; and encryption standards. ASSISTments invites users, parents/guardians, and educators who want more detailed documentation to request it.

Principle #6: Give users control over their data.

In addition to ensuring that students and others understand how their data is being used, it is critical to give them (or in the case of young users, their parents/guardians or teachers) control over their data. Users should be able to make informed decisions about how they interact with AI-enabled tools, particularly regarding what data they share.

Developers should:

- **Ensure informed consent.** Make users aware of what personal data the tool collects and which data, if any, may be shared for research or commercial purposes. Give users a clear and easy way to opt in or out of sharing that data at any time. In the case of partnerships with schools, it is the schools that "own" the student data and decide what gets shared with the developer. Do not collect data for which consent has not been given.

- **Disclose research participation and allow opting out.** For IRB-approved research that involves the tool, provide clear information to the schools, teachers, and parents/guardians of students who might participate in the research. Explain what participation entails and give them the ability to opt out.

- **Enable users to make choices about the collection and deletion of their data.** Users should have the ability to set their own data privacy preferences, namely what data they wish to share. Additionally, users (or their parents/guardians) should be informed of how to contact the developer to request the removal and deletion of their data. Ideally, allow users to export and retain their data.

- **Empower school staff with data insights.** Educators and administrators should have access to meaningful, privacy-protected data insights generated by the tool. This lets them to monitor how students are interacting with the AI-powered tool and, as needed, make adjustments to instruction.

What does this look like in practice?

GoalLight, the latest tool from EdLight (a 2025 Tools Competition winner), uses AI to support learning for students with an IEP. With AI, it interprets a student's handwritten math work and maps it directly to their learning objectives.

GoalLight shares with students, teachers, and families the data it collects. It provides them with personalized feedback and next steps aligned with the learner's unique plan. In particular, GoalLight empowers teachers by giving them research-backed analyses of student work, generating actionable insights, and offering ongoing, co-designed professional learning and coaching.

EdLight helps users control their data by linking to its user-friendly privacy policy on their website. The data policy clearly outlines what data users agree to share, the business rationale behind it, and which third parties may access it. These descriptions are tailored for different user types (e.g., student, parent, teacher). Additionally, EdLight helps institutional users review all the data policies during their onboarding process.

Building a Responsible AI Future

Learners, educators, and families deserve safe and responsible AI in education. By following these principles and checklists, developers can create AI solutions that are not only innovative but also ethical and effective. As AI continues to transform the educational landscape, developers have an enduring responsibility to prioritize their users' safety, privacy, and learning outcomes.

Other Frameworks

- AI Risk Management Framework (National Institute of Standards and Technology). Released in 2023, the U.S. Department of Commerce developed this detailed framework in collaboration with public and private stakeholders to help individuals, organizations, and society manage the risks of AI.

- Data Governance for EdTech: Summary of Landscape Review and Recommendation (UNICEF) In partnership with the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Global Privacy Assembly (GPA), UNICEF developed this review of the ed tech landscape and set of broad recommendations for safeguarding student data.

- Education Technology Industry's Principles for the Future of AI in Education (SIIA). This framework is intended to guide the implementation of AI technologies purposefully, transparently, and equitably. It incorporates principles around addressing user needs, protecting student data, and supporting AI literacy for students and educators.

- SAFE Benchmarks (EDSAFE AI Alliance). This guidance distills 24 global AI safety, trust, and market frameworks into four key principles: safety, accountability, fairness and transparency, and efficacy.

- Safety by Design (Australian Government). This framework is intended to center safety in the development of online and digital technologies. It emphasizes service provider responsibility, user empowerment and autonomy, transparency, and accountability.

- The Ethical Frameworks for AI in Education (The Institute for Ethical AI in Education). This detailed framework was developed in collaboration with a wide range of stakeholders. It reinforces the importance of equitable AI systems, user autonomy, data privacy, transparency, accountability, and ethical design.

- Recommendation on the Ethics of Artificial Intelligence (UNESCO). While this resource is focused on policy recommendations to promote the ethical use of AI, it includes the principles that guide them, including safety, fairness, sustainability, data protection, human oversight, transparency, and explainability.

Acknowledgments

We would like to thank the individuals who generously contributed to this report. We are grateful for the expertise and insights you shared with us.

Ralph Abboud
Principal Scientist, AI for Math and Education, Renaissance Philanthropy

Anna Aldric
Director of Technology Innovation, Axim Collaborative

Thomas Christie
Director, Engineering Hub, Renaissance Philanthropy

Kristen DiCerbo
Chief Learning Officer, Khan Academy

Jennie Dougherty
Director of Strategic Initiatives, KIPP Public Schools Northern California

Cristina Heffernan
Co-Executive Director and Founder, ASSISTments Foundation

Luis Pérez, PhD
Senior Director of Disability and Accessibility, CAST

Caitlin Mills, PhD
Associate Professor, University of Minnesota and Chief Research and Impact Officer, AugmentED

Jean Rhodes
Co-Founder, MentorPRO and Frank L. Boyden Professor of Psychology and Director of the Center for Evidence-Based Mentoring at UMass Boston

Hannah Horne-Robinson
Research & Assessment Manager, Rising Academies

Medha Tare, PhD
Senior Research Director, Joan Ganz Cooney Center at Sesame Workshop

Michel Meneses
Head of Engineering, EdLight

Nona Ullman
CEO, LessonLoop

References

Arciniega, J., Sexton, M., & Vance, A. “The K-12 Privacy Policy Guide: How to Quickly Spot Red Flags” (April 2024). https://publicinterestprivacy.org/privacy-policy-red-flags/

Burns, M., Winthrop, R., Luther, N., Venetis, E., & Karim, R. “A New Direction for Students in an AI World: Prosper, Prepare, Protect” (January 2026). https://www.brookings.edu/articles/a-new-direction-for-students-in-an-ai-world-prosper-prepare-protect/

EDSAFE AI Alliance. “SAFE Benchmarks” (2021). https://www.edsafeai.org/safe

eSafety Commissioner. “Safety by Design Overview” (May 2019). https://www.esafety.gov.au/sites/default/files/2019-10/SBD%20-%20Overview%20May19.pdf

Fesler, L., Martinez Claeys, J., Agnew, C., & Loeb, S. “The Evidence Base on AI in K-12: A 2026 Review” (March 2026). https://scale.stanford.edu/sites/default/files/The%20Evidence%20Base%20on%20AI%20in%20K-12%20Report.pdf

Learning Engineering Virtual Institute. “2026 Safeguarding Survey” (Administered in April 2026).

Sikka, M. “Why the Human-in-the-Loop Model is Key to Ethical AI in K-12 Education” (Accessed in April 2026). https://blog.definedlearning.com/why-the-human-in-the-loop-model-is-key-to-ethical-ai-in-k-12-education/

Software & Information Industry Association (SIIA). “Education Technology Industry’s Principles for the Future of AI in Education” (October 24, 2023). https://edtechprinciples.com/principles-for-ai-in-education/

The Learning Accelerator (now FullScale). “Driving EdTech Systems: Student Data Privacy” (Accessed in April 2026). https://practices.learningaccelerator.org/strategies/driving-edtech-systems-student-data-privacy

The Institute for Ethical AI in Education. “The Ethical Frameworks for AI in Education” (March 2021). https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

Tools Competition. “Safeguarding Edtech Users & Data Privacy” (November 20, 2024). https://tools-competition.org/safeguarding-edtech/

Tools Competition. “Webinar: Safeguarding Edtech Users & Data Privacy” (December 4, 2024).

Uncapher, M., Fitzgerald, B., Vance, A., Arciniega, J., Sexton, M., Sanberg, I., Parks, A., Osoba, T., & Whitmer, Ja. “Open Source Guidebook on Advanced R&D in Education” (August 5, 2025). https://osf.io/6dxgh/files/m3rz5

UNICEF Innocenti – Global Office of Research and Foresight. “Data Governance for EdTech: Summary of Landscape Review and Recommendation” (September, 2025). https://www.unicef.org/innocenti/media/11616/file/UNICEF-Innocenti-Data-Governance-Education-Technology-Summary-2025.pdf

United Nations Educational, Scientific and Cultural Organization (UNESCO). “Recommendation on the Ethics of Artificial Intelligence” (2022). https://unesdoc.unesco.org/ark:/48223/pf0000381137

Appendix

Developers' Checklist

Use this checklist to promote the safe and responsible use of AI in your ed tech tool.

Principle #1: Focus on educational outcomes.

- [ ] Rely on research and evidence.
- [ ] Co-design with users.
- [ ] Stay engaged with users.
- [ ] Ensure that the tool complements and enhances, rather than replaces or conflicts with, teacher instruction.
- [ ] For student-facing tools, put pedagogical guardrails in place to mitigate the risks of AI-enabled cognitive offloading.
- [ ] Center humans in implementation.
- [ ] Continuously measure and work to improve the accuracy rate of the tool's core AI task.
- [ ] Commit to evaluation and continuous improvement.

Principle #2: Direct generative AI systems to support students' learning and wellbeing.

- [ ] Train and routinely test the AI model to prevent its outputs from causing harm.
- [ ] Engage in red teaming.
- [ ] Put a content moderation plan in place to detect and respond to harmful AI outputs, as well as inappropriate inputs from the user.

Developers' Checklist (Cont.)

- [ ] Use Retrieval-Augmented Generation (RAG) and thoughtful prompt engineering to minimize hallucinations and biases.
- [ ] Invite users to flag issues.

Principle #3: Ensure the privacy and security of user data.

- [ ] Understand and adhere to privacy laws and regulations.
- [ ] Develop or strengthen data privacy policies.
- [ ] Collect data purposefully.
- [ ] Store data securely.
- [ ] Develop and execute a robust incident response plan for foreseeable threats to data security.
- [ ] Share data responsibly.
- [ ] Remove PII from any data used in a generative AI model.
- [ ] Engage in anonymization.
- [ ] Undergo data management audits.
- [ ] Offer educators data privacy support and training.

Principle #4: Prioritize accessibility and fairness.

- [ ] Consider and design for the diversity of the population the tool serves.
- [ ] Monitor for bias after initial design by tracking outputs.
- [ ] Promote accessibility for all students.

Developers' Checklist (Cont.)

- [ ] Test for compatibility with assistive technologies.
- [ ] Undergo accessibility audits.

Principle #5: Promote transparency and explainability.

- [ ] Articulate the tool's theory of change.
- [ ] Communicate policies and practices.
- [ ] Disclose to the user any tool features that use AI.
- [ ] Strive to make the tool's underlying AI system understandable to users.
- [ ] Make security and privacy documentation available to users.
- [ ] Keep users informed.

Principle #6: Give users control over their data.

- [ ] Ensure informed consent.
- [ ] Disclose research participation and allow opting out.
- [ ] Enable users to make choices about the collection and deletion of their data.
- [ ] Empower school staff with data insights.