EDUCATE VENTURES RESEARCH

In partnership with AVALLAIN

FROM THE GROUND UP

DEVELOPING STANDARD ETHICAL GUIDELINES FOR AI IMPLEMENTATION IN EDUCATION

Full Version

FROM THE GROUND UP

Developing Standard Ethical Guidelines

for AI Implementation in Education

EDUCATE VENTURES RESEARCH

AVALLAIN

WE HELP YOU REALISE

THE OPPORTUNITY OF AI

Foreword

From our beginnings, in the early days of digital education, the team at Avallain has been dedicated to developing educator-led technology. We believe a research-driven approach is essential to ensuring that technology is never an end in itself but a true enabler of learning. We have seen this in the evolution of online learning, mobile learning and adaptive learning. However, with the rapid rise of generative AI (GenAI), this approach is more important than ever.

Together with Educate Ventures Research (EVR), we aimed to explore how GenAI ought to be implemented into educational technology, not only to mitigate its risks but, more crucially, to support teaching and learning practices. We are now pleased to share the results of this research with the educational community at a time when clear guidelines and recommendations are essential. As we navigate the evolving field of AI, our goal remains steadfast: Deliver trustworthy technology that empowers and protects both teachers and students.

- Ursula Suter and Ignatz Heinz, Co-Founders of Avallain

At Educate Ventures Research, our mission has always been to bridge the gap between educational practice and technological innovation, ensuring that advancements in AI serve the needs of learners, teachers, and society as a whole. As the capabilities of AI rapidly evolve, so too must our frameworks for ethical and effective implementation. This is especially true in education, where trust, well-being, and human development are paramount.

In response to this need, EVR led a collaborative effort to develop a set of practical, research-informed ethical controls for the use of AI in educational technology. This framework was shaped by a diverse range of voices—including educators, learners, policymakers, parents, and developers—each of whom brought a vital perspective on what it means to implement AI responsibly in learning environments.

These twelve ethical controls are not abstract principles. They are grounded in the realities of the classroom and designed to support genuine teaching and learning. From preserving student agency and promoting critical thinking to ensuring cultural inclusion and safeguarding wellbeing, the framework helps EdTech providers and education leaders alike align innovation with integrity.

We view this work not as a conclusion, but as a foundation. The ethical use of AI in education is a living conversation—one that must evolve alongside the technologies we build and the societies we serve. We hope this report empowers schools and developers to take their next step with confidence, clarity, and care.

- Professor Rose Luckin, CEO, Educate Ventures Research

Summary

Introduction and Purpose

The rapid integration of artificial intelligence in educational settings offers promising opportunities for personalised learning, administrative efficiency, and expanded access to resources. However, this technological evolution brings significant ethical considerations that must be addressed to ensure AI serves educational values rather than undermining them. This report presents a comprehensive framework of twelve ethical controls for AI implementation in education, developed through rigorous research and extensive stakeholder consultation.

Ethics in education, particularly concerning AI implementation, is paramount due to the profound impact educational experiences have on shaping individuals and society. As AI systems become more prevalent in educational settings, they influence critical aspects of learning, assessment, and educational decision-making. Without proper ethical guidelines, there is a considerable risk of perpetuating biases, compromising student autonomy, or prioritising efficiency over holistic development. Moreover, as education plays a crucial role in moulding future citizens, the ethical use of AI in this domain sets a precedent for how society at large will interact with and govern these technologies.

Development Process and Methodology

The framework was developed through a multi-phase process over approximately six months in 2024-2025. The approach combined systematic literature review, case study analysis, and extensive stakeholder engagement to ensure the guidelines would be both theoretically sound and practically applicable.

The initial phase involved foundational research, including a systematic literature review of existing AI ethics guidelines and analysis of case studies where educational institutions or adjacent sectors implemented AI controls. By late 2024, researchers formulated an initial draft of AI controls addressing identified gaps in current guidance.

Stakeholder consultation was central to the development process. A teacher focus group reviewed the draft framework, contributing practical insights about classroom realities. Concurrently, a multidisciplinary expert panel was convened, comprising educators, school administrators, AI ethicists, and edtech industry specialists. This panel participated in workshops using the Delphi process, providing structured feedback on each proposed control. The project also drew on concerns and findings from a landscape study conducted by Educate Ventures Research, which engaged educational leaders from 23 multi-academy trusts encompassing 413 schools and 250,000 students.

The framework underwent multiple iterations based on stakeholder input. Teacher feedback led to clearer definitions and practical use-case examples for each control. The expert workshop resulted in the expansion of existing controls and the addition of two new ones. Throughout the consultations, the team continuously refined the framework, integrating stakeholder suggestions and resolving ambiguities to create a comprehensive and actionable set of guidelines.

The Twelve Ethical Controls

The framework consists of twelve ethical controls, each addressing a distinct aspect of AI implementation in educational settings:

1. Learning Outcome Alignment: Ensures AI tools continuously support the full spectrum of learning goals rather than narrow metrics. This control requires implementing a continuous evaluation system for AI-driven educational interventions that assesses both immediate academic outcomes and long-term educational impact across diverse learning objectives.

2. User Agency Preservation: Designs AI systems to empower users with choice and control, ensuring that AI in education does not undermine student autonomy or teacher professional judgment. The AI should act as a supportive guide rather than an autocratic tutor, with safeguards against over-automation.

3. Cultural Sensitivity and Inclusion: Ensures AI educational tools are culturally responsive and free from bias, providing an inclusive experience for all learners. This entails establishing systematic processes to detect and correct cultural biases in AI content or interactions, with diverse representation in training data and knowledge bases.

4. Critical Thinking Promotion: Embeds opportunities for students to practice critical thinking when using AI-powered tools. Rather than passive acceptance of AI-generated outputs, the system should prompt reflection and scepticism, encouraging students to question, analyse, and critically evaluate information.

5. Transparent AI Limitations: Clearly communicates what the AI can and cannot do to all stakeholders. This control implements user-friendly explanations about AI systems' capabilities, limitations, and decision processes to manage expectations and prevent misplaced trust.

6. Adaptive Human Interaction Balance: Maintains a healthy balance between AI-mediated learning and human-human interaction. Guidelines establish thresholds for minimum human engagement, ensuring that AI personalisation does not come at the expense of essential teacher-student and peer-to-peer interactions.

7. Impact Measurement Framework: Establishes a framework to measure the real educational impact of AI interventions, both short-term and long-term. This combines quantitative data with qualitative assessments in regular review cycles to gauge how AI affects learning and inform improvements.

8. Ethical Use Training and Awareness: Provides mandatory training for all stakeholders on the ethical and appropriate use of AI in education. These programs cover topics such as academic integrity, understanding AI bias, privacy issues, and responsible use policies, tailored to different stakeholder groups.

9. Bias Detection and Fairness Assurance: Implements continuous processes to detect, audit, and mitigate bias in AI systems to ensure fair educational opportunities for all students. This includes using specific fairness metrics, conducting regular audits, and establishing clear processes for addressing identified biases.

10. Emotional Intelligence and Well-being Safeguards: Monitors and supports student emotional well-being in AI-mediated learning, with protocols for human intervention when needed. This control balances emotion detection with privacy and non-intrusiveness, establishing protocols for human response when an AI detects possible emotional issues.

11. Organisational Accountability & Governance: Establishes institutional oversight and clear lines of responsibility for AI systems used in education. This control creates governance frameworks—policies, committees, and processes—to ensure AI tools are deployed ethically and in compliance with legal requirements.

12. Age-Appropriate & Safe Implementation: Ensures that AI tools and practices in education are tailored to students' developmental stages and uphold a safe, child-friendly learning environment. This includes configuring content and capabilities suitable for different age groups, implementing content filtering, and prioritising child safety and well-being.

Table of Contents

Introduction & Approach

Ethics in education, particularly concerning AI implementation, is of paramount importance due to the profound impact educational experiences have on shaping individuals and society. As AI systems become more prevalent in educational settings, they influence critical aspects of learning, assessment, and educational decision-making. Ethical considerations are essential to ensure that these technologies are deployed in ways that respect student privacy, promote fairness and inclusivity, and uphold the fundamental values of education. Without proper ethical guidelines, there's a risk of perpetuating biases, compromising student autonomy, or prioritising efficiency over holistic development. Moreover, as education plays a crucial role in moulding future citizens, the ethical use of AI in this domain sets a precedent for how society at large will interact with and govern these technologies. Therefore, establishing robust ethical frameworks for AI in education is not just about protecting students; it's about shaping an ethically conscious, technologically adept future society.

Development Process

The framework was developed through a multi-phase process over approximately six months in 2024-2025. The approach combined a systematic literature review, analysis of case studies, and extensive stakeholder engagement to ensure the guidelines would be both theoretically sound and practically applicable.

The initial phase involved foundational research, including a systematic literature review of existing AI ethics guidelines and analysis of case studies where educational institutions or adjacent sectors implemented AI controls. By late 2024, researchers formulated an initial draft of AI controls addressing identified gaps in current guidance.

Stakeholder consultation was central to the development process. A teacher focus group reviewed the draft framework, contributing practical insights about classroom realities. In parallel, we convened a multidisciplinary expert panel, comprising educators, school administrators, AI ethicists, and edtech industry specialists. This panel was chosen keeping in mind not only their expertise and relevance to the field of AI and education, but also representation from across geographies - so as to include the widest possible diversity of opinion. This panel participated in workshops using the Delphi process, providing structured feedback on each proposed control.

The project also drew on concerns and findings from a landscape study conducted by Educate Ventures Research, which engaged educational leaders from 23 multi-academy trusts encompassing 413 schools and 250,000 students. This wide stakeholder engagement helped ensure the framework addresses real-world needs across students, teachers, parents, educational institutions, and edtech developers.

The guidelines went through multiple iterations. Early teacher feedback led to clearer definitions and use-case examples for each control, making them more actionable. After the expert workshop, certain controls were expanded or added to cover gaps the experts noted. For instance, the need for explicit Organisational Accountability mechanisms and Age-appropriate safeguards emerged strongly and resulted in the addition of two new controls in those areas. Definitions were refined to incorporate expert-recommended language – for example, emphasising "balanced innovation with oversight" in the accountability control, drawing on expert advice that AI should enhance, not replace

human roles. Throughout consultations, the team revisited the framework, integrating stakeholder suggestions and resolving ambiguities.

The Importance of Ethical Controls for AI in Education

As artificial intelligence becomes increasingly integrated into educational settings, it is essential to ensure its implementation aligns with core educational values and ethical principles. AI has the potential to personalise learning, streamline administrative tasks, and expand access to resources. However, without ethical safeguards, these advancements risk compromising student privacy, reinforcing biases, or prioritising efficiency over meaningful learning experiences.

Ethical controls serve as a framework to guide responsible AI use in education, ensuring that technology supports—not replaces—human-centered learning. Students, particularly younger learners, may not always recognise when AI-driven decisions impact their education. Without proper oversight, AI could inadvertently shape learning trajectories in ways that limit opportunities, reinforce stereotypes, or introduce unfair advantages. Similarly, teachers and educational institutions require clear ethical guidelines to balance AI's capabilities with professional judgment, fostering trust and accountability in AI-enhanced learning environments.

Beyond immediate classroom implications, ethical AI use in education sets a precedent for how future generations will engage with technology. Schools play a critical role in shaping digital citizens who will navigate an increasingly AI-driven society. By embedding ethical principles into AI adoption today, we ensure that education remains equitable, transparent, and aligned with the broader mission of developing informed, capable, and ethically aware individuals.

Complete List of Refined AI Guidelines

Following is the list of the 12 guidelines for ethical AI implementation in education. Each guideline is presented with a definition, challenges (including stakeholder-specific difficulties), mitigation strategies, implementation guidance (drawing on best practices from research and other sectors), and its relevance to key stakeholder groups (students, teachers, parents, institutions, developers).

1. Learning Outcome Alignment

Definition

Ensure AI tools continuously support the full spectrum of learning goals. This control requires implementing a continuous evaluation system for AI-driven educational interventions that assesses both immediate academic outcomes and long-term educational impact. AI recommendations and actions should be regularly checked against diverse learning objectives – not just test scores, but also higher-order thinking, creativity, engagement, and skill development over time. The system should track traditional metrics alongside qualitative indicators (e.g. critical thinking, collaboration) and adjust AI behaviour to keep alignment with curriculum goals and student development needs. Clear roles must be defined for educators, developers, and policymakers in reviewing outcome data and refining the AI system's approach.

Challenges

Different stakeholders face distinct challenges in aligning AI with holistic learning outcomes. For students, a narrow AI focus (for example, drilling test answers) can neglect broader skills, reducing opportunities for creative or critical thinking. They may not immediately perceive the importance of certain AI-driven tasks that target long-term skills. Teachers might struggle if the AI's recommendations emphasise easily measurable outcomes (like quiz scores) at the expense of harder-to-measure goals like socio-emotional growth or creativity – this can conflict with teachers' professional judgment about well-rounded education. Parents may worry that an AI tutor or curriculum tool is "teaching to the test" and not fostering 21st-century skills or values. Educational institutions find it challenging to evaluate AI impact beyond exam performance; long-term studies are resource-intensive, and schools may lack frameworks to measure outcomes like critical thinking or learner autonomy. EdTech developers encounter technical difficulties in designing AI systems that optimise for multi-dimensional outcomes – it's far easier to optimise for a single metric (e.g. accuracy on exercises) than for a broad set of cognitive and developmental goals. Additionally, proving long-term efficacy (e.g. over several school years) is difficult during development cycles.

Mitigation Strategies

- Develop a comprehensive framework of learning outcomes that includes both quantitative measures (grades, test scores) and qualitative indicators (student curiosity, collaboration, creativity). This multi-faceted rubric should guide AI behaviour and evaluations.
- Conduct regular reviews of AI recommendations by educational experts or curriculum specialists to ensure they align with broad educational objectives, not just narrow targets. If the AI's suggestions start drifting toward a limited set of skills, human reviewers can intervene to recalibrate the system.
- Design the AI to suggest a balanced mix of learning activities addressing various outcomes. For instance, for each AI-recommended practice quiz, it might also recommend an open-ended project or discussion prompt. This ensures that even if the AI optimises for certain outcomes, it still engages students in diverse learning modes.
- Implement a multi-year impact assessment plan to track students' progress beyond immediate AI interactions. This could involve following up on how students who used the AI perform in subsequent grades or in qualitative aspects like their confidence and independence as learners.
- Program the AI with the ability to adjust its objectives per student once certain short-term goals are met. For example, if a student has mastered factual recall, the AI can shift focus to higher-order applications of knowledge.

Implementation Guidance

Academic research underscores the importance of long-term and holistic evaluation of AI in education. Studies have found that positive effects of educational technology often only emerge over extended use – for example, a large multi-school trial showed no significant improvement from an AI algebra tutor in the first year, but significant gains appeared in the second year of implementation (Pane et al., 2013). This suggests schools should commit to sustained use and evaluation, rather than expecting instant results. Regular longitudinal studies and feedback loops are vital; one systematic review found it "unclear… how [AI] can actually impact meaningfully on teaching and learning," noting a lack of evidence for long-term benefits and a need for more robust evaluation methods (Zawacki-Richter et al., 2019). To address this, educational institutions can borrow best practices from healthcare and business: in healthcare, new interventions are subject to multi-year trials and outcomes tracking, and in corporate training, programs are evaluated on both immediate performance and long-term employee growth. Applying similar rigour, schools should establish cross-functional committees involving educators, data analysts, and learning scientists to periodically review AI impact data. Integrating insights from learning science is also recommended – for instance, frameworks for evaluating learning impact (Luckin & Cukurova, 2019) advocate combining test results with observations of student meta-cognition and engagement. Additionally, leveraging policy guidance can help - the "Shape of the Future" (2024) education leaders' report urges developing comprehensive evaluation frameworks that consider both short-term and long-term impacts of AI on learning. Schools might partner with researchers to design assessments that capture skills like critical thinking or creativity fostered (or hindered) by the AI. In practice, implementing this control may involve using dashboards that show a variety of student performance indicators, not just one score, and training teachers to interpret these reports.

Relevance to Stakeholders

Students: Ensures the AI tools they use contribute to meaningful learning, not just higher test scores. This control protects students from being funnelled into shallow learning and helps them build a richer skill set (critical thinking, creativity) which benefits their long-term success.

Teachers: Supports teachers by aligning AI assistance with curriculum goals and pedagogy. It prevents conflict between what the AI pushes and what teachers know is important, thus allowing teachers to trust and use AI as a complementary tool. When the AI accounts for higher-order outcomes, teachers can more readily integrate its suggestions into lesson plans.

Parents: Reassures parents that AI in the classroom or at home is not a "black box" obsessing over grades, but is monitored for educational value and well-rounded development. It addresses parental concerns that technology might shortchange their child's broader learning (like creativity or social skills) by explicitly keeping those outcomes in focus.

Educational Institutions: Aligns AI implementations with school or district educational missions (e.g. producing creative, well-rounded learners, not just high test performers). It provides institutions with accountability – through documented evaluation cycles – to demonstrate that adopting AI is improving educational quality in a broad sense. It also helps in accreditation or compliance contexts by showing that tech use meets learning standards and strategic goals.

EdTech Developers: Guides developers to design products that measure and report a variety of learning outcomes. It pushes them to go beyond one-dimensional success criteria, potentially giving their product a competitive edge in efficacy. However, it also means developers must invest in educational research and perhaps collaborate with educators to define and embed these multi-faceted metrics into their AI systems.

2. User Agency Preservation

Definition

Design AI systems to empower users with choice and control. This control ensures that AI in education does not undermine student autonomy or teacher professional judgment. AI systems should provide meaningful options for students, teachers, and even parents to shape the learning process – for example, allowing students to set personal learning goals, choose between different types of learning activities, or override AI recommendations when they see fit. The AI should act as a supportive guide rather than an autocratic tutor. Safeguards must be implemented to prevent over-automation: in other words, the AI should never take away all decision-making. The design ethos is that the user (learner or educator) remains the ultimate decision-maker on the learning journey.

Challenges

The primary challenge is avoiding scenarios where users become overly reliant on AI guidance. Students may begin to follow AI suggestions uncritically, clicking whatever the tutor system or educational app recommends, thus failing to develop self-directed learning skills. This "automation complacency" can diminish their agency and critical thinking. Conversely, if given choices, some students might feel overwhelmed or make suboptimal decisions, especially younger learners – striking the right balance of guidance vs. freedom is hard. Teachers might struggle to integrate student choices with curriculum requirements; they may fear that giving students too much say (facilitated by AI) could lead to off-track learning or classroom management issues. Teachers also need to preserve their own agency – if an AI grading system or content recommendation engine is too rigid, teachers might feel they must follow it even when it contradicts their expertise. Parents could find it difficult to trust an AI-driven system if they feel it either disempowers their child or conversely, gives their child too much leeway in a way that might reduce academic rigour. Institutions must balance standardised instruction with personalised pathways; giving every student a custom path can complicate scheduling, assessment, and ensuring coverage of standards. Developers face the challenge of designing AI interfaces that invite user input at key junctures without compromising the AI's effectiveness. They also have to guard against users making choices that render the AI less useful (for instance, a student consistently opting out of challenging tasks the AI recommends).

Mitigation Strategies

- Introduce deliberate "choice points" in the AI workflow where students (or teachers) must make active decisions. For example, after an AI presents a learning module, it could ask the student to choose one of two projects to apply the knowledge. This prevents passive following and exercises the learner's decision-making muscles.
- Implement a "learning reflection" feature that occasionally prompts students to justify or reflect on their choices and on the AI's suggestions. For instance, if a student sticks with the easiest quizzes, the system might ask, "Why did you choose this path? Would you like to try a harder challenge?" Such prompts encourage metacognition and ensure the student's agency is coupled with responsibility.
- Calibrate the level of autonomy according to the learner's development. One mitigation is to gradually increase the number of student-directed choices as they gain proficiency. A novice might start with a more guided experience; as they demonstrate

capability, the AI offers more open-ended options. This approach, akin to scaffolding, balances guidance and independence.

- Ensure teachers have override authority at all times. If the AI recommends a particular content sequence or flags a student for intervention, the teacher should be able to adjust or countermand that based on their professional judgment. Clear teacher dashboards can allow easy modifications of AI-prescribed learning plans.
- Include brief in-app tutorials or tips educating users (students and teachers) on how to effectively use the AI as a tool they direct, rather than as an oracle. This could mitigate over-reliance by framing the AI as one source of input among many. For instance, the system might remind users: "These suggestions are here to help, but feel free to explore and choose what's best for you."

Implementation Guidance

Preserving agency aligns with established best practices in human-computer interaction and education psychology. Research shows that learners benefit from a sense of control over their learning pace and path, which enhances motivation and engagement. However, too much choice without guidance can lead to decision fatigue or aimless exploration. To implement this control, draw on user-centred design principles: involve students and teachers in co-designing the AI system's interface and choice architecture. In fields like personalised learning and intelligent tutoring, studies have found that when students can set goals or choose help levels, they develop better self-regulation skills. At the same time, system designers have used techniques like fading guidance – gradually removing supports as competence grows – which is effective in both educational software and professional training. Another strategy gleaned from other sectors (e.g. aviation or medicine) is the concept of human-in-the-loop: no critical decision is made by AI alone. In educational AI, this means always having a human checkpoint (student or teacher) before significant changes, similar to how autopilot systems still require pilot input at critical moments. Notably, Eaton (2023) describes the emerging "post-plagiarism" era where students use AI assistance for schoolwork and argues that education must adapt by explicitly teaching students to wield AI tools responsibly rather than banning them. This underscores the need for training learners in making informed choices with AI – essentially nurturing agency. The Delphi panel of experts in our project strongly emphasised maintaining a "healthy balance between AI assistance and user autonomy" as a guiding design principle. In practice, developers might implement this by including toggle settings (e.g. a student can switch a recommendation mode on or off) or multi-path lesson plans. Importantly, allow personalisation: some learners will want more AI guidance, others less, so systems should be flexible to individual agency preferences.

Relevance to Stakeholders

Students: Keeps students in the driver's seat of their learning. They benefit by developing decision-making and self-regulation skills. This control also helps maintain their motivation and identity as learners – rather than feeling controlled by an algorithm, they see the AI as a helpful resource that they direct, leading to more meaningful engagement.

Teachers: Respects teachers' professional agency by ensuring the AI does not override their decisions. Teachers can trust that they can use the AI as a supportive tool without losing control over their classroom or lesson planning. It essentially positions AI as an aide that teachers can configure, rather than an instructor that dictates to them.

Parents: Provides assurance that their child isn't just blindly following a computer's commands. Parents can appreciate that the system is designed to teach their children how to make good choices and become independent learners. It also means parents can be involved – e.g. the AI might allow a parent to set certain learning goals or preferences for their child, giving families a voice in the process.

Educational Institutions: Aligns with educational values around student-centred learning and personalised education. Schools and districts are increasingly prioritising student agency (for example, inquiry-based learning models); this control ensures that adopting AI won't run counter to those initiatives. It also helps with compliance to any standards that emphasise student choice or differentiation.

EdTech Developers: Encourages the creation of AI products that are empowering rather than restrictive. This can improve user satisfaction – students and teachers are likely to favour tools that they feel in charge of. However, it also challenges developers to design intuitive interfaces for choice and to avoid algorithmic "lock-in" where the AI's way is the only way. In the long run, products that successfully preserve user agency can become preferred solutions in education, as they will integrate more smoothly into varied teaching styles and student needs.

3. Cultural Sensitivity and Inclusion

Definition

Ensure AI educational tools are culturally responsive and free from bias, providing an inclusive experience for all learners. This control entails establishing systematic processes (like audits and feedback loops) to detect and correct cultural biases in AI content or interactions. AI systems should be built and maintained with diverse representation in training data and knowledge bases, so that examples, language, and context are appropriate for learners from different backgrounds. Regular "cultural audits" of the AI's outputs (whether it's curriculum suggestions, automated feedback, or even examples used in problems) must be conducted. The AI should incorporate multiple cultural perspectives and be adaptable to local educational values – for example, a story problem might have regionally relevant names or scenarios. Inclusion also means ensuring the AI is accessible and welcoming to learners of various abilities, languages, and identities.

Challenges

AI systems can inadvertently propagate cultural bias or insensitivity. A common challenge is that many AI models are trained on data predominantly from Western or other dominant cultures, leading to outputs that may be irrelevant or even offensive in other cultural contexts. Students from underrepresented groups might find that the AI "doesn't speak to them" – e.g. examples that assume certain holidays, lifestyles, or idioms unfamiliar to them, which can disengage or alienate. In worst cases, biases can manifest as stereotypes (say, an AI career advisor suggesting different roles to male vs. female students due to biased training data). Teachers face the challenge of spotting and addressing subtle biases from the AI; they may not always notice if an AI's feedback consistently favours one group of students or if content is skewed, especially if it's in areas outside the teacher's own background. Parents, especially from minority communities, might distrust AI tools if they see them reflecting a monocultural or biased viewpoint, leading to reluctance in adoption. Institutions must ensure compliance with equity and anti-discrimination policies – deploying an AI that ends up biased could lead to public relation issues or even violations of equity regulations. They also face the logistic challenge of customising AI for each local context (a national or global product might not fit their community's culture out-of-the-box). Developers often struggle to obtain sufficiently diverse training datasets and to recruit experts from multiple cultures to evaluate AI behaviour. It's also challenging to define metrics for "cultural inclusivity" – unlike clear accuracy metrics, inclusivity is qualitative and broad, making testing and validation difficult.

Mitigation Strategies

- Use globally representative training datasets and knowledge sources. During development, explicitly include data (texts, examples, user scenarios) from a wide range of cultures, ethnic groups, and locales. This reduces the chance that the AI's default behaviour is culturally one-sided.
- Conduct regular cultural audits of the AI system's outputs. This means periodically reviewing lesson content, feedback, and any AI-generated material for cultural bias or blind spots. Ideally, form an advisory board with members from various cultural backgrounds to perform these reviews and provide guidance. Such a group can flag content that a homogeneous team might overlook.

- Implement an easy-to-use feedback/flagging system where students and teachers can flag content that they find culturally insensitive or not inclusive. For example, if a student notices an AI-generated example that they find offensive or exclusionary, they can click “Flag this” – and such flags trigger review and improvements. This crowdsources cultural vigilance to the users experiencing the AI in real contexts.
- Design the AI to be locally customisable. For instance, allow educators to input local context info (like local holidays, names common in their community, relevant cultural references) so the AI can adapt scenarios accordingly. Balance global and local content by letting schools toggle certain content sets or upload their own culturally relevant materials that the AI can incorporate.
- Before deployment, test the AI for bias across different demographics. For example, run the AI’s essay scoring or tutoring feedback on inputs that simulate various dialects or cultural content to see if it treats them equitably. Use metrics like content diversity (does the AI’s output equally include references from different cultures?) and performance parity (does the AI perform as well for a student in one country as another?). Engage external auditors or use bias-detection tools to evaluate the AI regularly.

Implementation Guidance

This control is informed by extensive literature on AI bias and the need for inclusivity. Recent research highlights issues like representation bias – AI reflecting the cultural biases present in training data. Chinta et al. (2024) note that addressing this requires more than just diversifying data; it needs active involvement of stakeholders from different cultures throughout development. In implementing this control, a useful model comes from the concept of “AI ethics of care”, which suggests continuous engagement with the communities impacted. For example, Singapore’s national education AI framework places strong emphasis on cultural sensitivity: they conduct regular stakeholder engagements and even cultural audits of AI systems to ensure alignment with local values. Following that example, educational institutions might hold community forums or student focus groups about the AI’s content and behaviour, treating feedback as a critical component of AI governance. Another relevant concept is “AI colonialism” – the idea that AI tools developed in one context (often Western, English-speaking) could impose values or patterns on other cultures. Researchers have warned that without local adaptation, AI in education could inadvertently perpetuate cultural inequities. Thus, implementation should include contextual adaptation: customising AI to fit into the cultural norms of each educational setting, rather than a one-size-fits-all deployment. Technical guidance can be drawn from the NIST AI risk management framework and similar guidelines which emphasise fairness and bias mitigation as core tenets (National Institute of Standards and Technology, 2023). These frameworks call for involving subject-matter experts to evaluate AI systems in context and for ongoing monitoring – exactly what cultural audits and diverse governance boards would do. In practice, developers can use tools like inclusive design checklists (many organisations provide AI bias checklists) and integrate libraries or models that support multilingual and multicultural content. It’s also advisable to pilot the AI in a small, diverse set of classrooms initially, observe its interactions, and refine before scaling up.

Relevance to Stakeholders

Students: Provides a more inclusive and relatable learning experience. Students from all cultural backgrounds should see themselves reflected in the learning material and not feel marginalised by examples or content that assume a different norm. This increases engagement and a sense of belonging. Importantly, it protects students from the harm of stereotyping or bias – e.g., ensuring an AI mentor encourages equally high expectations for all students, regardless of background.

Teachers: Helps teachers by providing AI tools that are attuned to the classroom's cultural makeup. Teachers won't have to constantly "translate" or contextualise AI-provided content to make it relevant – saving time and avoiding potential classroom missteps. Moreover, it supports teachers in upholding equity; they can trust that the AI is not introducing bias that they then need to undo. If issues do arise, the feedback mechanisms empower teachers to get them corrected.

Parents: Increases parent trust in educational AI systems. Parents are more likely to support AI use if they see it respects their culture and values. For instance, a parent who notices the AI giving assignments that honour their cultural heritage or at least don't demean it will feel more comfortable. In communities with historical educational disparities, demonstrating this commitment to inclusion can be critical to adoption.

Educational Institutions: Meets schools' obligations to provide equitable education. Many districts have diversity and inclusion policies – using an AI that has built-in cultural sensitivity aligns with those policies and reduces the risk of biased outcomes (like achievement gaps widening because the AI only resonated with some students). It also enhances the institution's image as progressive and culturally competent. Should any issues be identified, the institution's process of auditing and addressing them (as this control entails) can serve as evidence of proactive governance.

EdTech Developers: Following this guideline makes products more globally marketable and socially responsible. While it adds development overhead, it opens products to wider audiences (multi-language, multi-regional use) and can prevent harms that might lead to backlash or regulatory action. Developers also benefit from user feedback systems – by learning from flags and audits, they can improve the product continuously. There is also a growing expectation (from regulators and customers) that AI products demonstrate fairness and bias mitigation; adhering to this control helps developers meet standards such as the IEEE guidelines on algorithmic bias or upcoming AI regulations that require bias risk assessments.

4. Critical Thinking Promotion

Definition

Embed opportunities for students to practice critical thinking when using AI-powered tools. This control requires that AI in education should not only deliver content or answers but also actively encourage students to question, analyse, and critically evaluate AI-generated outputs. Instead of learners passively accepting recommendations or answers from an AI tutor/assistant, the system should prompt reflection and scepticism. For example, an educational AI might provide an answer and then ask the student, "Do you think this answer is correct? Why or why not?" or present an alternative perspective to discuss. Features like structured reflection prompts, guided debates, or self-assessment questions are integrated throughout the AI-supported learning experience. The goal is to use the presence of AI as a springboard to deepen students' critical thinking, so they learn to not take information at face value – even (and especially) when it comes from an AI.

Challenges

Without intentional design, AI tools can inadvertently promote passivity. Students often perceive computer-provided answers as authoritative; if an AI tutor gives a solution, many will accept it without question, which can weaken their habit of critical analysis. Some may also become dependent on AI explanations and not attempt to solve or reason through problems themselves ("why think hard if the AI will tell me the answer?"). Teachers face the challenge of ensuring that the use of AI doesn't short-circuit the learning process. For instance, if an AI homework helper provides a worked solution, a student might copy it; the teacher then has to gauge whether the student actually understands the material. It's challenging to design assignments or class activities that leverage AI while still requiring students to think. Parents might be concerned that easy access to answers (via AI) undermines the development of grit and problem-solving in their children. They worry about a "calculator effect" but for reasoning – i.e., kids not learning to reason because an AI spoon-feeds them. Institutions need to uphold academic integrity and rigour; integrating AI in learning must not dilute the development of higher-order thinking outcomes that curricula and standards emphasise. There's also the risk of misinformation – if students are not critical, AI-generated content (which might occasionally be incorrect or biased) could mislead them. Developers must figure out how to keep students engaged in thinking when an AI could just give the result. It can be non-trivial to program an AI to intentionally hold back or inject questions without frustrating users who "just want the answer." Additionally, measuring whether an AI tool is successfully fostering critical thinking is difficult – traditional metrics might not capture that.

Mitigation Strategies

- Integrate "AI scepticism" exercises into the platform. For instance, occasionally the AI could present a statement and ask the student to find flaws or verify it through research. This could also mean allowing students to challenge AI responses (such as in the case of feedback). These exercises train students to scrutinise AI outputs.
- Include built-in prompts for reflection whenever the AI provides information. The system might ask, "Can you explain this in your own words?" or "What might be a counterargument or alternative answer?" prompting the learner to process and evaluate the AI's contribution actively.

- Encourage peer discussion and debate about AI-provided content. For example, in a classroom setting, an AI system could provide two different approaches to a problem to two groups of students and then prompt a debate on which is better. Collaborative tools where students critique or build on AI outputs together can reinforce critical evaluation.
- Design assessments that require students to apply or critique AI outputs. A mitigation for over-reliance is if assignments explicitly ask students to analyse an AI's answer – e.g., "The AI gave this essay a high score. Do you agree with this evaluation? Provide your own critique of the essay." This way, even if AI is used, the student's critical role is mandatory.
- Educate users that the AI is not infallible. The interface can gently remind, "AI can be wrong – always double-check and think for yourself." If the AI occasionally makes a known benign error and then guides the student to spot it, that could be a powerful learning moment. Essentially, use the AI's imperfections as cases for critical thinking (of course, without compromising core learning).

Implementation Guidance

The importance of teaching students to critically evaluate AI outputs is increasingly recognised in academic literature and practice. As generative AI and other systems become prevalent, educators are shifting focus from preventing AI use to embedding AI literacy – which includes critical thinking about AI. A practical example comes from many schools (especially in higher education) that now incorporate lessons on identifying AI-generated text and checking it for accuracy, instead of outright banning it. Our framework draws on such emerging best practices. Research by Montenegro-Rueda et al. (2023) suggests that explicit instruction on evaluating AI content can improve students' analytical skills and awareness of AI limitations. One approach is to treat the AI as a "cognitive apprentice" model: have students critique the AI's reasoning just as they would a peer's reasoning in class. In fields like media literacy, educators have students dissect news articles for bias; similarly, in AI literacy, students can dissect AI responses for correctness and bias. Tools and curriculum are already being developed for AI literacy in K-12 – for example, some schools use scenarios where an AI provides conflicting answers and students must determine which (if any) is correct, thereby practicing evidence-based reasoning. It's also recommended to incorporate this into teacher training: teachers should be trained in strategies to prompt and guide critical thinking when AI is part of learning (such as always asking "How did the AI get that? Do we trust this source?" in classroom discussions). Notably, the European Network for Academic Integrity has highlighted that rather than trying to eliminate AI from student use, it's more practical to teach students how to use it critically and ethically. In implementation, developers should consider features like a "challenge mode" where the AI intentionally leaves a gap or poses a question for the student to fill in. Additionally, institutional policies could encourage that at least X% of assignments with AI involvement require students to submit a reflection on how they used the AI and what they learned from it. The Shape of the Future report (2024) noted many schools have begun explicitly teaching AI bias and misinformation as part of the curriculum, which is a supportive measure for this control.

Relevance to Stakeholders

Students: Equips students with a crucial 21st-century skill: the ability to think critically about AI and information in general. This means students become less likely to be misled by wrong answers and more adept at reasoning. In the long run, it prepares them to interact intelligently with AI in higher education or the workplace, always adding human judgment to AI output. It also makes learning more engaging – instead of just being told an answer, they get to play detective or evaluator, which can deepen understanding.

Teachers: Aligns with teachers' goals of developing students' critical thinking and not just content mastery. It provides teachers with tools and prompts to spark discussions about why an answer is correct or not, facilitating richer classroom dialogue. By building these checks into the AI, it reduces the burden on teachers to constantly interject "don't just copy that" – the system itself promotes the behaviour. Teachers also gain more confidence that if students are using AI, they are still learning the "how and why," not just the "what."

Parents: Addresses parental concerns about AI making students intellectually lazy. Parents can be shown that the AI is designed to ask students "why do you think that?" and similar questions, thus actually strengthening their child's thinking skills. This can make parents more supportive of AI tools, seeing them as a way to foster independent thought, not stifle it.

Educational Institutions: Ensures that introducing AI does not compromise educational quality or integrity. Schools can maintain that their graduates still meet critical thinking learning outcomes. In fact, schools can boast that they are teaching a new kind of critical thinking – not just traditional argument analysis, but also scepticism of AI and automation – which is increasingly important. It helps institutions fulfil any critical thinking components of curricula and can be a metric to showcase in accreditation reviews (e.g. demonstrating how technology integration still supports higher-order thinking objectives).

EdTech Developers: By focusing on critical thinking, developers differentiate their products as pedagogically sound and not just efficient "answer machines." This can appeal to educators and administrators making adoption decisions. It may require more sophisticated AI (capable of generating explanations, alternative answers, etc.), but it aligns the product with ethical education values. Developers also mitigate risk – a platform that actively teaches users to verify its answers is less likely to be responsible for serious misinformation going unchallenged. Essentially, users won't blame the AI for being wrong if the AI itself asked them to double-check. Over time, such design could become a standard expectation (akin to how calculators have a display to show steps, etc.), so early adoption is forward-looking.

5. Transparent AI Limitations

Definition

Clearly communicate what the AI can and cannot do to all stakeholders. This control ensures transparency about AI systems' capabilities, limitations, and decision processes in an educational context. Practically, it means implementing user-friendly explanations and disclosures whenever an AI is in use. For example, an AI writing assistant might have a pop-up or info box saying, "I am a language model trained on text and might make errors. Verify the facts I provide." Similarly, a dashboard for teachers might label recommendations with confidence levels or notes about the source of the AI's suggestion. The idea is to manage expectations and prevent "misplaced trust" by making sure students and educators understand the scope and reliability of the AI. Transparency features can include real-time explanations (showing a summary of why the AI suggested something), periodic reminder messages about the AI's purpose and limits, and thorough documentation or help menus detailing the AI's training data, bias mitigation, and appropriate use cases.

Challenges

A key difficulty is communicating technical information in an age-appropriate and context-appropriate manner. Students (especially younger ones) may not grasp what it means that "the AI is not 100% accurate" or the concept of AI training data biases. If transparency isn't done well, warnings might be ignored or misunderstood. On the other hand, too much transparency (or technical detail) can overwhelm or confuse users. Teachers and parents need enough information to trust the system but not so much that it deters usage. For instance, a teacher might see a low confidence score and then doubt whether to ever use the AI's recommendation – balancing trust and caution is tricky. Another challenge is that transparency windows or prompts could interrupt the user experience; students might find them annoying and dismiss them without reading. Institutions must ensure that transparency doesn't become merely a box-ticking compliance item (like a long terms-of-service nobody reads). They need it to be effective in informing consent and usage. There's also the challenge of keeping transparency up-to-date: as the AI gets updates or new features, the information given to users must be revised accordingly. Developers face the technical challenge of generating explanations for AI decisions (e.g., why did the AI flag this essay for potential plagiarism? Why is it suggesting this lesson to this student?). Many AI models (like deep neural nets) are complex and not easily interpretable. Developers also might worry that exposing limitations could reduce user confidence or reveal proprietary model info. Finally, achieving transparency across different literacy levels is tough – the messaging might need to be different for a 4th-grade student versus a high schooler versus a teacher.

Mitigation Strategies

- Develop clear, age-tailored explanations of AI capabilities and limitations. For younger students, this might be a simple cartoon or analogy (e.g., "I'm still learning, so I might mess up sometimes!"). For older users, provide slightly more detail (e.g., "This AI was trained on data up to 2022, so it might not know recent events.").
- Implement periodic "AI awareness" pop-ups or notifications within the learning platform. For example, if a student uses the AI helper for an hour continuously, a gentle reminder can appear: "Remember to double-check answers – AI can make mistakes." These reinforce awareness without relying on them to read a manual.

- Create a short orientation module or guide for teachers and parents about the AI. For instance, a 10-minute online course or a documentation packet can be provided, explaining in non-technical terms how the AI works, its benefits, and its limitations. Ensuring parents understand, for example, that an AI tutor doesn't replace human teaching or that it might not get context perfect, will manage expectations at home.
- Use visual design to communicate confidence or limits. One mitigation is showing a confidence meter or different colour coding when the AI is less certain. If an AI answer is given with low confidence, it might appear with a warning icon. If it's a type of problem the AI hasn't seen much (a limitation), maybe a small note "experimental suggestion" is attached. Such cues quickly tell users how much caution to exercise.
- Encourage and enable users to ask "Why?" For example, alongside an AI's suggestion, have a button: "Why did you suggest this?" Clicking it would show a brief rationale (even if simplistic, like "I noticed you struggled with fractions, so I suggested more practice on that."). This not only informs the user but also engages their critical thinking (tying in with the previous control). Over time, this can build appropriate trust: users learn the AI's patterns and limits through these micro-explanations.

Implementation Guidance

Transparency is a core principle in Responsible AI frameworks globally. The European Union and many academic bodies stress transparency to ensure users are not misled by AI. A concrete set of recommendations comes from Foltýnek et al. (2023) via the European Network for Academic Integrity, which "strongly emphasises the importance of transparency regarding AI capabilities and limitations", noting that misunderstandings of AI can lead to misuse. They advocate clear communication to maintain proper expectations. In practice, several educational platforms have started implementing transparency features. For example, some AI-based language learning apps display messages like "Generated by AI" or provide sources when giving factual info. These can be emulated. It's also useful to take cues from sectors like healthcare: medical AI tools often include disclaimers ("This is not a medical diagnosis") and require a human professional to interpret results – similarly, educational AI can say "Not a grade: teacher makes final grading decisions" for an AI grader, as an example. The NIST AI Risk Management Framework (2023) lists transparency as one of the pillars of trustworthy AI, suggesting organisations document and communicate AI limitations as part of risk mitigation (National Institute of Standards and Technology, 2023). Our implementation should include maintaining a transparent documentation portal for the AI – a place where curious stakeholders can read about how it works, what data it uses, and known limitations or past incidents. Additionally, survey and research within the institution can guide how to refine transparency messaging: gather input from students and teachers on whether they understand the AI's role and any misconceptions. One finding from the "Shape of the Future" report was that many students and educators overestimate AI capabilities, sometimes expecting near-human intelligence, which can lead to over-reliance. The report underscores providing clear protocols to communicate AI limitations to all stakeholders as essential. Therefore, training sessions at the rollout of an AI (like a webinar for teachers or an assembly for students) can set the tone by openly discussing what the AI can/can't do. Finally, always accompany transparency with humility: encourage a culture where it's fine to point out when the AI is wrong – this complements the critical thinking control and reinforces the understanding of limitations.

Relevance to Stakeholders

Students: Helps students develop a correct mental model of the AI tutor or tool. They learn not to blindly trust it and to treat it as a helpful aide with potential flaws. This protects them from learning incorrect information and teaches an important lesson about technology – that it has limits. It can also reduce frustration: if students know an AI might not understand a poorly worded question, they won't be as frustrated when it fails, because they expected some limits.

Teachers: Empowers teachers with knowledge about the AI, increasing their confidence in using it. When teachers understand exactly what the AI is doing behind the scenes and its known limitations, they can better integrate it into instruction (and know when to step in). Transparency ensures teachers maintain authority in the classroom – since they and their students know the AI is a tool with constraints, the teacher remains the ultimate authority on content.

Parents: Builds trust with parents by demonstrating that the school is not deploying a mysterious "black box" on their child. Instead, parents see that the school is being upfront about the AI's role and limitations. For example, a parent portal might show a note, "Your child used an AI reading coach today – it helps with pronunciation but might not catch all nuances, so we encourage you to also listen to them read." Such transparency invites parents to partake in oversight, which they appreciate.

Educational Institutions: Satisfies ethical and potentially legal requirements for informed consent and transparency in technology use. Many education systems require notifying if student data is used by an AI or if automated decisions are made. This control ensures institutions are proactively doing so, reducing risk of backlash or non-compliance. It also fosters a culture of openness, which is valuable for community relations – schools can say to their community, "We are using AI, and here's exactly how it works and what it does," pre-empting fear or rumours.

EdTech Developers: Although developers might fear that highlighting limitations could reduce user confidence, in the long run it increases trust by preventing disillusionment. Users who understand limitations won't have unrealistic expectations and then be upset when the AI fails at something it was never meant to do. Moreover, developers often get blamed for AI mistakes; clear disclaimers and explanations can mitigate liability and support proper use (users are less likely to misuse an AI if they know its boundaries). Regulators, too, are increasingly likely to mandate transparency features – by implementing them, developers stay ahead of regulation and demonstrate responsible innovation, which can be a market differentiator.

6. Adaptive Human Interaction Balance

Definition

Maintain a healthy balance between AI-mediated learning and human-human interaction. This control sets guidelines to ensure that the introduction of AI personalisation or tutoring does not come at the expense of essential teacher-student and peer-to-peer interactions. It involves establishing thresholds or norms for minimum human engagement. For instance, a school might decide that even with AI-assisted practice, every student must have at least one substantial teacher-led discussion per day, or that group work should occupy a certain percentage of class time. Technologically, it means designing AI systems to complement rather than replace human interaction. The AI could, for example, alert a teacher when it detects a student might benefit from one-on-one help (flagging possible disengagement or confusion). Features might include automatic reminders to take breaks from the computer to discuss or collaborate with others. The overarching goal is to guard the social dimension of education – recognising that communication, empathy, and social skills are built through human interaction, which must be preserved even as AI use grows.

Challenges

One challenge is identifying the right "balance" – how much human interaction is enough? Students vary; some may thrive with more independent, AI-guided study, while others desperately need social learning. Rigid thresholds might not fit all situations. Also, if an AI is very engaging or effective, there's a temptation (by both students and teachers) to lean on it heavily, possibly inadvertently reducing human interaction. Teachers might struggle to integrate AI sessions and traditional interactions fluidly – e.g., managing class time so that an AI activity transitions into a group discussion requires planning and skill. There is also the risk that teachers might rely on AI for certain tasks (like answering student questions) and unintentionally become less available to students. Parents could be concerned if they see their child spending long hours isolated on an AI platform. Especially after periods of remote learning, many parents and experts worry about excessive screen time and lack of socialisation. Institutions face logistical challenges: ensuring classes still incorporate enough face-to-face engagement might mean limiting AI usage, which could conflict with technology adoption goals or be hard to enforce. They also need criteria to measure interaction quality, not just quantity. Developers might not initially prioritise features to encourage human interaction, since the AI's aim is often to engage the learner with the software itself. It takes conscious design to include things like "Now discuss with your classmate" prompts, which might even reduce time spent in-app (a disincentive for some edtech business models). Additionally, detecting when a student needs human help (to trigger an alert) can be complex and prone to error.

Mitigation Strategies

- Set and monitor minimum thresholds for human-led instruction and collaboration. For example, in a blended class, mandate that at least 30% of class time is discussion or group work. If using an AI tutor at home, perhaps for every X minutes on the AI, the student should explain what they learned to a family member or in writing (simulating human reflection). By policy or design, ensure a baseline of human contact.
- Design AI systems to identify opportunities for human interaction and actively prompt them. For instance, if a student has been working solo with the AI for a long stretch, the system could suggest, "Take a break and explain what you've learned to a friend or

teacher." If a student is struggling repeatedly, it might pop up, "Consider asking your teacher for help on this topic." These prompts serve as built-in reminders to involve humans at key moments.

- Implement features that facilitate easy transitions between AI and human-led activities. For example, the AI platform could have a "Group Mode" where it provides a discussion question that a small group can tackle together, effectively handing off to peer learning. Or a teacher dashboard might allow the teacher to see where students are in the AI exercise and then pause the AI for the whole class to have a live discussion about a common challenge that came up. Designing interoperability between AI tools and classroom routines helps blend the two modalities.
- Use the AI's monitoring capabilities to flag when a student might be disengaging or isolated. For example, if the AI notices a student rapidly clicking through without learning (potentially frustrated), it could alert the teacher: "Student X might need a check-in." Also, if a student hasn't interacted with peers on an assignment (while others have), the system might remind the teacher or student to incorporate some collaboration. By acting as a support system for teachers, AI can ensure no student "falls through the cracks" of human attention.
- Build in structured checkpoints where human interaction is required. For instance, an AI homework system might require that each student's session summary is reviewed by a teacher or discussed in class the next day. This ensures the AI use is not in isolation – it's always bracketed by human feedback or discussion.

Implementation Guidance

Studies in education and HCI repeatedly show that blended approaches (AI + teacher) outperform AI-alone or teacher-alone in many cases, due to the complementary strengths of each. Holstein et al. (2020) emphasise designing for human-AI complementarity, arguing that AI should augment rather than replace human teachers, with research evidence that the best outcomes occur when AI is used to support rich human instructional interactions. Implementers should heed such findings by intentionally keeping teachers in the loop and ensuring the AI is seen as a tool, not a tutor in isolation. In practice, some innovative schools have already instituted policies like "face-to-face Fridays" or similar to ensure human connection in tech-rich environments. Our control might inspire guidelines such as: if an AI is providing practice exercises, the teacher will still do the conceptual introduction and subsequent discussion of misconceptions – the AI's role is confined to practice. Another piece of guidance comes from developmental psychology: children (especially younger ones) require social interaction for healthy cognitive and emotional development. Over-reliance on AI for learning could reduce those critical interpersonal experiences. Thus, an age-sensitive approach may apply: the younger the student, the more stringent the human interaction requirement (perhaps tying in with the Age-appropriate Implementation control). The "Shape of the Future" report (2024) noted that education leaders were concerned about displacement of valuable human-to-human learning experiences and emphasised maintaining meaningful human interaction as AI use grows. They recommended clear thresholds for teacher-student engagement and ongoing monitoring of interaction quality, aligning directly with this control. For implementation, it could be useful to leverage classroom observation frameworks: administrators or instructional coaches can observe how AI is used in class and note the ratio of tech vs human dialogue, helping teachers adjust if needed. Technology can also assist; for example, some platforms provide analytics on how often teachers intervene or students collaborate

on the platform, giving quantitative measures of interaction. Ultimately, maintain a pedagogical design where AI is woven into lesson plans that also include discussions, group tasks, and teacher-led inquiry. Training teachers on blended learning strategies is key – so they know how to orchestrate activities to achieve this balance.

Relevance to Stakeholders

Students: Preserves the social aspects of learning that are crucial for engagement, motivation, and holistic skill development. Students continue to benefit from mentorship, empathy, and dynamic discussion with teachers and peers, even as they use AI tools. This balance helps them develop teamwork and communication skills alongside individual learning. It also protects against feelings of isolation – students don't end up learning in a silo with a machine but remain part of a learning community.

Teachers: Ensures teachers remain central to the learning process and their roles are valued. Teachers get to focus on what humans do best – inspiring, guiding, and addressing emotional and higher-level needs – while the AI handles repetitive practice or data. This can improve job satisfaction and effectiveness, as teachers spend less time drilling and more time interacting meaningfully with students. By receiving AI-driven alerts or data, teachers can intervene at the right time, making their interactions more targeted and impactful.

Parents: Addresses parental concerns about excessive screen time and losing the human touch in education. Parents can be assured that, even with AI, their child will still be working with teachers and classmates frequently. Many parents view school as an important place for social development; this control guarantees that technology will not upend that. It also means parents might still hear their child talk about their teacher or friends, not just about a computer program, when discussing what they learned – a reassuring sign of balanced development.

Educational Institutions: Supports well-rounded educational outcomes (including social-emotional learning) and aligns with mandates to provide a safe, collaborative learning environment. Schools uphold their educational philosophy that values relationships and interaction. It can also serve as a metric of quality: institutions might track and report that "even with AI integration, student engagement with teachers and peers remained high." Additionally, it may alleviate any pushback from those who fear AI will "replace teachers" – clearly, with this control, the institution's stance is that teachers stay irreplaceable.

EdTech Developers: Encourages developers to create features that integrate with classroom workflows instead of trying to monopolise student attention. While it might reduce time-in-app, it can increase the educational effectiveness and adoption of their product because schools and teachers will prefer tools that respect pedagogical balance. Also, by focusing on complementarity, developers can carve a niche where their AI fills specific gaps and explicitly hands off other parts to humans, making it easier to market as a teacher's partner rather than a teacher replacement. It also opens opportunities for new features – like teacher alert systems or collaboration modes – which can differentiate their products in a positive way.

7. Impact Measurement Framework

Definition

Establish a robust framework to measure the real educational impact of AI interventions, both short-term and long-term. This control calls for a comprehensive evaluation methodology that goes beyond simple metrics and captures the holistic value (or drawbacks) of AI in the educational environment. It involves combining quantitative data (test scores, completion rates, engagement analytics) with qualitative assessments (teacher feedback, student surveys, observations of classroom dynamics) to gauge how AI is affecting learning. The framework should include national or regional benchmarking where applicable (to see if AI-using classrooms perform better or differently than others) and compliance indicators if relevant (e.g., ensuring AI use aligns with standards or policies). Importantly, it should be an ongoing process: regular review cycles (say each semester or year) are established to assess effectiveness and inform improvements. In essence, this control institutionalises the practice of treating AI implementations as evidence-based initiatives that require data-driven validation.

Challenges

Measuring impact is complex for several reasons. Attribution is one: if student outcomes improve or decline, it's hard to attribute how much is due to the AI versus other factors (teacher skill, curriculum changes, socio-economic factors, etc.). Timescale is another challenge – some benefits or harms of AI might only manifest in the long run (e.g., critical thinking skills might erode or improve over years, not weeks). Many schools operate on short evaluation cycles, which might miss long-term effects. Data collection difficulties also arise when qualitative data (like student attitudes or teacher perceptions) takes effort to gather and analyse. Quantitative data might be easier (e.g., the AI can log performance continuously), but making sense of it (are higher quiz scores translating to better understanding?) is tricky. Teachers and staff may feel burdened by yet another layer of assessment to conduct. They might also be biased in reporting (e.g., a teacher who has a positive view of AI might overestimate its impact in surveys). Institutions might face pressure to show positive impact (especially if money was invested in AI), which could bias analyses or create reluctance to acknowledge negative findings. Parents could be sceptical of how impact is measured ("are test scores all that matter?"). And developers may not have an immediate incentive to facilitate rigorous independent evaluations of their products, unless required – meaning schools might have to undertake impact studies largely on their own or with third parties.

Mitigation Strategies

- Develop a multi-year assessment plan to track outcomes beyond the initial implementation. Plan to follow cohorts of students for several years, comparing those who extensively use the AI versus those who don't, to capture long-term effects. Built-in checkpoints (e.g., end-of-year exams, next-grade readiness indicators) should be included.
- Use diverse evaluation methods. Quantitative: pre-and-post implementation test scores, assignment grades, learning analytics (time on task, error rates, etc.). Qualitative: interviews or focus groups with teachers and students, classroom observations focusing on engagement or behaviour changes. By having multiple data sources, you can triangulate the true impact. For instance, improved standardised test

scores (quantitative) accompanied by teacher observations of deeper project work (qualitative) provides strong evidence of positive impact.

- Include assessments that capture real-world skill application, not just grades. For example, project-based evaluations or portfolios that show how students apply knowledge could be used to see if the AI's support translates into better skill application. If the AI claims to improve critical thinking, have students do a performance task that requires critical thinking and have it evaluated by independent educators.
- Where possible, collaborate with researchers to design quasi-experimental evaluations. If you can't randomly assign AI vs non-AI (often not feasible), consider comparing similar classes or schools with and without the AI, or doing a phased rollout where late adopters serve as a comparison in the interim. Engaging educational researchers can lend rigour and help control for confounding variables.
- Make the measurement framework participatory. Regularly present interim findings to teachers, students, and parents and get their feedback on whether it matches their experience. This can reveal impacts that numbers don't show or identify additional outcomes to measure (for instance, if students say, "I feel more confident now," that could be added as an outcome measure via surveys). It also ensures transparency – stakeholders see that the school is critically evaluating the AI, not just blindly using it.

Implementation Guidance

The impetus for this control comes from the recognition that despite enthusiasm for AI in education, solid evidence of its effectiveness is limited and mixed. Zawacki-Richter et al. (2019) in their review stressed that many claims about AI benefits lack rigorous backing and that more systematic evaluation is needed (Systematic review of research on artificial intelligence applications in higher education – where are the educators? - UCL Discovery). Our framework acknowledges that and requires institutions to treat AI implementations almost like educational interventions or pilot programs that need evaluation. One can draw on methodologies from program evaluation and improvement science in education: set clear goals (e.g., "improve maths problem-solving by 15% in one year"), then monitor progress and iterate. Also, consider the breadth of impact: Al-Zahrani (2024) suggests including social and emotional development in evaluating AI's impact, not just academic metrics. This means the framework might include measures of student well-being or collaboration levels in AI-augmented classes. The "Shape of the Future" report (2024) highlights that system leaders want comprehensive evaluation of both immediate and long-term impacts, and it suggests regular evaluation cycles with feedback from all stakeholders – essentially exactly what this control is about. They also emphasise considering whether AI tools actually improve learning outcomes versus just making processes efficient. A practical tip is to align the AI impact measures with existing metrics the school cares about – for example, if a school already measures reading levels thrice a year, see if the AI is moving the needle on those. Also incorporate any national exam results if applicable but be careful to account for curriculum alignment. Many sectors, like business and healthcare, use Key Performance Indicators (KPIs) and continuous improvement cycles; education can adopt a similar approach for AI: define KPIs for AI (e.g., "reduce homework non-completion by X%" or "increase student self-efficacy scores by Y"), monitor them, and adjust strategies accordingly. It's wise to publish or at least document the findings formally – it could contribute to broader knowledge on AI in education.

On the tech side, developers should provide analytics and data export tools to help schools track usage and outcomes. Over time, if done well, the measurement framework will illuminate what works and what doesn't, allowing the school to refine how the AI is used or decide if it's worth continuing. If the data shows little to no benefit, that's a sign the AI might need to be modified or even shelved in favour of other solutions – a tough but important decision that this control forces in the name of evidence-based practice.on the platform, giving quantitative measures of interaction. Ultimately, maintain a pedagogical design where AI is woven into lesson plans that also include discussions, group tasks, and teacher-led inquiry. Training teachers on blended learning strategies is key – so they know how to orchestrate activities to achieve this balance.

Relevance to Stakeholders

Students: Ultimately benefits students by ensuring that the AI tools being used are actually helping them learn. It prevents them from being subject to ineffective or harmful tools for long periods. If something isn't working (say the AI isn't actually improving their understanding), the measurement framework will catch it and prompt changes. It also can highlight positive impacts, which can then be communicated to students (e.g., "Since using this tool, the class as a whole is writing longer, more complex essays!"), which can boost morale and buy-in.

Teachers: Involves teachers in a reflective process about the AI's role. Teachers gain insight from the data – maybe they see that the AI really helped with factual recall, so they can adjust their teaching to focus more on higher-order skills, for example. If the framework includes teacher feedback, their professional observations are valued and can lead to support or adjustments (like more training if impact is lacking). Also, if positive impact is demonstrated, it validates teachers' efforts in adopting and learning the new technology. If negative or neutral, it ensures teachers aren't forced to continue with something that doesn't work, freeing them to try other methods.

Parents: Gives parents concrete information on how the AI is affecting their child's education, countering uncertainty. Instead of vague assurances, schools can share evidence ("In classes using the AI tutor, test scores improved by 10% on average (Does an Algebra Course with Tutoring Software Improve Student Learning? | RAND), and no adverse effects on homework quality were observed"). This transparency can build trust. Moreover, parents want to know that instructional time is used effectively – this framework shows that the school is vigilant about ROI on learning. If the data showed negative effects (like too much screen time hurting grades), parents would want the school to know and act; this control ensures that happens.

Educational Institutions: Enables school leaders to make data-informed decisions about technology investments and pedagogy. It supports a cycle of continuous improvement – aligning with approaches like PDCA (Plan-Do-Check-Act) in educational leadership. The institution can demonstrate accountability: to school boards or education authorities, they can provide reports on AI implementation outcomes, showing responsible stewardship. Over time, this can contribute to research and the school's reputation as a thoughtful, evidence-based innovator. It also mitigates risk: if the AI is causing issues, early detection allows course correction before any major damage (academic or reputational).

EdTech Developers: While independent evaluations can be a double-edged sword for developers, in the long run they push developers to improve their products. Constructive feedback from measurement (especially if schools share it) can highlight areas where the AI falls short or excels. Developers who partner with schools on impact studies may gain credibility – positive results can be published as case studies (with appropriate rigour), aiding marketing. If results are negative, it's a chance to iterate on the product. Additionally, if many schools use similar frameworks, developers might start anticipating these needs by building in analytics and conducting their own efficacy research to align with school expectations. In a broader sense, an industry norm of proven impact could emerge, rewarding those companies that genuinely enhance learning.

8. Ethical Use Training and Awareness

Definition

Provide mandatory training for all stakeholders on the ethical and appropriate use of AI in education. This control involves developing and delivering comprehensive education programs about AI ethics, tailored to different groups – students, teachers, and possibly parents. The training should cover topics such as academic integrity when using AI (e.g., avoiding plagiarism with AI help), understanding AI bias, privacy issues, and how to use AI tools responsibly and in alignment with school policies. It should use practical scenarios and case studies to illustrate potential ethical dilemmas and best-practice responses. For example, a module for students might simulate discovering an AI-generated essay and ask how to handle it, or a module for teachers might present a scenario of a student relying too much on AI and how to intervene. Regular updates to this training are necessary since AI tech and norms evolve. Ultimately, the goal is to cultivate an informed school community that uses AI in a way that upholds academic values, equity, and safety.

Challenges

One challenge is keeping the training engaging and relevant – students (and adults) may tune out if it's too abstract or preachy. Students might see ethics training as an add-on not directly relevant to their immediate interests, so it must be made relatable. Also, some might have the attitude "I know how to use tech, I don't need this," especially digital-native teens. Teachers are very busy; finding time for thorough training and follow-ups can be hard. Additionally, teachers themselves are learning about AI capabilities – their personal comfort varies, so designing training that is neither too basic for some nor too advanced for others is tricky. Parents are another audience to consider; while they might not get formal training, raising their awareness (through workshops or communication) is part of this control – and reaching all parents, especially those less involved, is a challenge. Moreover, ethical AI use is not a one-time lesson – it requires a culture shift and continual reinforcement. Institutions need to invest resources into developing quality content and possibly bringing in experts. They also have to update these materials as AI tools change (for instance, the sudden appearance of a tool like ChatGPT requires rapid training updates). Ensuring consistency – that all teachers are enforcing the same ethical guidelines, that all students have a baseline understanding – can be hard in larger schools. Developers of edtech might not typically provide ethics guidance for using their tools (beyond terms of service), so schools can't rely on vendors for this; they must craft it themselves, potentially with expert help.

Mitigation Strategies

- Develop interactive, scenario-driven training modules that simulate real-life situations. For example, one scenario might be: a student is tempted to have an AI write an essay – what are the consequences and better choices? By walking students through these stories, they can better internalise the lessons. For teachers, scenarios might include detecting AI-generated work or addressing bias in an AI's recommendation. Make it case-study rich so it doesn't feel theoretical.

- Implement ongoing "ethics check-ins" – not just a one-off training. This could be brief discussions in class prompted by current events ("hey, an AI error made news, let's talk about it"), regular reminders of policies, or quick quizzes in home room. Continual dialogue keeps awareness high.

- Where possible, embed aspects of AI ethics into the curriculum itself. For example, in a digital citizenship or IT class, include a unit on AI. In social studies, one might discuss the societal impacts of AI. By making it part of the learning fabric, students see it as important and relevant, not an extra.
- Create easy-to-understand guidelines and tip-sheets for ethical AI use. For instance, a student-facing "Dos and Don'ts with AI in School" poster or infographic can reinforce training content. Likewise, for parents, send home a one-pager that summarises how the school encourages students to use (and not use) AI for homework, and how they can help at home.
- Host workshops or info sessions for the broader school community. Perhaps an evening seminar for parents and students together on "AI in Education – Using it Safely and Fairly," where school staff and maybe guest experts talk about these issues. This opens conversation and signals that the school treats this seriously. Some of this happened already with internet safety education; a similar model can be applied for AI.

Implementation Guidance

Many educational bodies are converging on the idea that AI literacy includes ethics. For example, UNESCO and other organisations have begun issuing guidelines on AI in education that emphasise training users in ethical and effective use, rather than just restricting use. In practice, some school districts have already started requiring academic integrity modules that specifically mention AI-assisted cheating and why it's wrong. Eaton (2023) notes the rise of "hidden" AI use by students and argues that proactive ethical guidance is critical. She suggests that it's better to educate students on appropriate use than to purely police them, aligning with our approach of training rather than just punishment. The "Shape of the Future" (2024) report found many schools implementing AI awareness and ethics training for staff and students; it even mentions hope for broader, possibly mandated AI literacy training in society. A best practice is to start the training early – as soon as AI tools are introduced, or even beforehand if possible – to set norms from the outset. Continuous PD (professional development) for teachers is also important: Samala et al. (2024) emphasise focusing on developing understanding of appropriate AI application rather than just listing don'ts. So, training should not be just "don't cheat" but also "here's how AI can be used beneficially" – for instance, teaching students to use AI for practice questions or brainstorming but not for final answers. This positive framing can make training more empowering. The European Network for Academic Integrity's recommendations, as cited earlier, effectively call for mandatory training programs for both educators and students on AI ethics – providing a strong external mandate to do exactly this. Implementation might involve collaboration with outside experts: perhaps partnering with a university or an organisation specialising in digital citizenship to develop the curriculum. Measuring the effectiveness of the training is also wise (quizzes, surveys on attitudes before and after). Lastly, keep the content updated: for example, if new AI tools emerge that can do novel things (like deepfakes or voice clones), update the training to address those, keeping everyone aware of current risks and responsibilities.

Relevance to Stakeholders

Students: Equips students with knowledge and values to navigate a world with AI. They learn how to use AI as a tool without violating academic integrity or compromising their learning. In effect, it helps them avoid potential pitfalls (like being accused of plagiarism or becoming too dependent on AI) by understanding the boundaries. It also empowers them – understanding AI better means they can use it more effectively and innovatively within the allowed limits. Ultimately, it contributes to their development as ethical digital citizens, a skill set that extends beyond school into higher education and careers.

Teachers: Provides teachers clarity and confidence on how to manage AI usage in their classes. Instead of each teacher having to figure out their own stance or disciplinary approach, the training and guidelines give a consistent framework. This makes enforcement of rules (like when AI help is allowed on homework) more straightforward and fairer. It also educates teachers themselves on AI capabilities and ethics, which helps them model proper use. Teachers, often being the first line of addressing issues like cheating, get concrete strategies from training on prevention and response, reducing anxiety about the unknowns of AI in student hands.

Parents: Reassures parents that the school is proactively guiding students on how to use new technologies responsibly. Many parents are themselves unsure about AI (some might not even know what tools their kids could be using), so the school taking initiative demystifies it and invites parents to be partners in reinforcing ethical use. Parents appreciate that their children are being taught not just academics but also values and decision-making. It also means fewer unpleasant surprises – like discovering their child used AI inappropriately – because the child has been taught clear rules and the importance of following them.

Educational Institutions: Helps maintain academic standards and a culture of integrity. By training everyone, the institution reduces incidents of misuse that could damage its reputation (e.g., widespread plagiarism scandals). It also aligns with legal and ethical obligations – for instance, educating minors about responsible technology use is increasingly seen as part of a school's duty of care (similar to internet safety training). Having a documented training program might also protect the institution: if issues occur, they can show they took reasonable steps to prevent them. Additionally, an informed community will likely handle AI transitions more smoothly and innovatively, potentially leading to better outcomes and easier implementation of new tech initiatives.

EdTech Developers: Indirectly, this fosters a more informed user base for developers' products. If students and teachers understand AI better, they may use the products more effectively and within intended use cases (e.g., not trying to force the AI to do something it shouldn't and then blaming it). For developers focusing on ethical design, a user community that values ethics will appreciate those design features (like an academic integrity mode). On a larger scale, widespread ethical AI training may lead to more trust in edtech, which can expand the market. It can also reduce negative uses of their tools (like if a tool was being used to cheat, and training curbs that, the tool's reputation remains positive). Some developers might collaborate by providing educational materials or built-in tutorials to support schools' efforts in this control, thereby showing corporate responsibility.

9. Bias Detection and Fairness Assurance

Definition

Implement continuous processes to detect, audit, and mitigate bias in AI systems to ensure fair educational opportunities for all students. This control mandates that schools and developers actively monitor AI tools for any form of bias – whether it's along lines of race, gender, language proficiency, disability, or other characteristics – that could lead to unequal outcomes. It includes using specific fairness metrics (such as checking that error rates or recommendations are equitable across student subgroups) and conducting regular audits of AI outputs and decisions. If an AI grading system is used, for example, this control would require analysing its scores to ensure one demographic isn't consistently scoring lower without justification. There should also be clearly defined responsibilities and processes for bias mitigation – i.e., if bias is found, who will address it and how (developers retraining model, adjusting algorithms, etc.). Correction mechanisms might include recalibrating the AI or instituting human review overrides when potential bias is detected. The goal is to uphold fairness, meaning the AI should neither disadvantage nor unfairly advantage any group of students.

Challenges

Bias can be hard to detect because it often requires large data analysis and understanding of context. A subtle bias (like the earlier example of an AI essay scorer giving lower scores to non-native speakers for language issues irrelevant to content knowledge) might go unnoticed without careful study. Data collection for bias analysis means gathering demographic or group information, which raises privacy and sensitivity issues – schools have to be careful about how they use such data. Teachers might not be trained in statistical methods to spot bias in AI outputs, so they may miss patterns. They might see individual odd cases but not realise a systemic issue. Institutions may not have data scientists on hand to do fairness checks, and small schools may lack sample sizes to assess bias meaningfully. There's also the challenge of defining fairness – for example, ensuring an AI gives equal opportunity might mean sometimes treating students differently (like providing more help to those with less prior knowledge) which complicates simple parity metrics. Developers may be reticent to share detailed model internals or involve external auditors due to intellectual property concerns. Also, fixing bias if found can be non-trivial (it might require collecting new data, which is expensive, or fundamentally changing algorithms). Students and parents might lose trust quickly if any bias incident occurs ("the AI is against my group"), making it urgent yet delicate to handle bias findings.

Mitigation Strategies

- Regularly audit AI outputs for bias patterns. For instance, each semester, review data like AI-assigned grades, suggestions, or flags broken down by student subgroups (gender, ethnicity, etc., as appropriate and legally/ethically permissible) to see if outcomes diverge. If, say, an AI maths tutor gives significantly more "struggle" flags for girls than boys, that warrants investigation. These audits can be done by an appointed committee or in partnership with an external evaluator to ensure objectivity.

- Employ specific fairness metrics to quantify bias. Examples: statistical parity (each group gets similar outcomes), equal opportunity (each group has similar success rates when they have similar inputs), or error rate balance (no group has a systematically higher error or failure rate from the AI). If using an AI for assessment, check inter-rater

reliability across groups (does the AI align with human graders equally for all groups?). Use these metrics in the audits and set thresholds that trigger action.

- Before deployment and periodically, run simulated test cases through the AI to detect bias. For example, create a set of assignments that are identical except for a student name that implies a certain ethnicity or gender; see if the AI responds differently. Or test an AI tutor with language input typical of English language learners vs. native speakers to see if it treats errors appropriately. This controlled testing can reveal biases in how the AI interprets different dialects or cultural references.
- Involve a diverse team in reviewing AI decisions. For instance, if an AI flags student essays for potential plagiarism or misconduct, have a panel of teachers from diverse backgrounds review those flags to check for any biased patterns (like perhaps it flags certain communication styles more often). A varied team is more likely to catch biases from their perspectives.
- Establish up-front how biases will be corrected when found. Perhaps set up an agreement with the AI vendor that if bias is detected, they will assist in retraining the model or adjusting parameters. Meanwhile, have interim fixes: e.g., if an AI grader is found biased, immediately switch to human double-grading for affected groups while the issue is addressed. Document these steps in an "AI fairness policy." By having this in place, when a bias is uncovered, everyone knows the plan (stop using the biased function or apply a correction factor, notify stakeholders, etc.).

Implementation Guidance

Fairness in AI is a heavily researched area, and education should borrow best practices from fields like finance or hiring where algorithmic bias has been tackled. For example, the finance industry uses adversarial testing to ensure credit models don't discriminate – similarly, educational organisations can adopt adversarial approaches to poke at their AI systems for weaknesses. Baker and Hawn (2021) provide many examples of AI bias in education and stress the necessity of robust detection and mitigation strategies. They also highlight that bias can creep in at various stages (data, algorithm, user interaction), which means our detection can't just be one-and-done; it needs to be continuous and multifaceted. Holstein et al. (2020) found that involving diverse stakeholders in the audit process is key, which backs our suggestion of diverse review teams. The NIST AI RMF and other guidelines often advise an "always on" monitoring of AI for performance and bias with feedback loops for improvement (Sullivan, 2023). That implies schools should treat bias checking as a routine maintenance task for AI, much like updating virus protection on computers – it's part of the ongoing operation. On the developer side, if vendors know that client schools are auditing their tools, they may proactively provide bias assessment results or tools (some may even provide an audit log or bias mitigation features built-in). Schools can pressure vendors by prioritising those who demonstrate fairness (perhaps ask in RFPs or procurement: "provide any evidence of fairness testing"). The Ogunleye et al. (2024) reference in the doc suggests ongoing monitoring and stakeholder consultation are vital to ensure fairness across different student populations, reinforcing that it's not a one-time checkbox but a continuous dialogue. It's also worth educating students and parents about what the school is doing to ensure fairness – this can build trust in the AI system. If an issue is found and corrected, transparently communicating that (with sensitivity to not erode confidence too much) can show the commitment to equity. Additionally, in some jurisdictions, algorithmic bias might have legal implications (anti-discrimination laws could apply if an AI systematically disadvantages a protected group), so this control also serves to keep the institution in compliance.

Relevance to Stakeholders

Students: Protects students from unfair treatment by AI systems. It ensures that no student is systematically shortchanged – for example, that a learning recommendation system gives all students the help they need, not just those from certain backgrounds. For students from historically marginalised groups, this control is critical for equity: it helps guarantee the AI won't inadvertently reinforce existing disparities. Overall, it upholds the principle that each student gets a fair chance to learn and succeed with the help of AI.

Teachers: Gives teachers confidence that the AI tools they use are equitable and alerts them to any potential issues. Teachers are often attuned to fairness in their classroom; this control extends that vigilance to the AI domain. If the AI is found to be biased, teachers can be part of the solution (e.g., adjusting how they use it or providing additional support to affected students). It prevents scenarios where teachers might unknowingly trust an AI that's disadvantaging some of their students. Instead, with audits and metrics, teachers get a clearer picture of the AI's behaviour across their diverse class and can respond appropriately.

Parents: Reassures parents that the school is actively ensuring the new technologies won't discriminate or treat their child unfairly. This is particularly important for parents of students who have unique needs or are from minority communities – they might be sceptical of an algorithm treating their child fairly. Knowing there are fairness checks and balances can ease concerns. If a bias is discovered and communicated, parents will at least know the issue is being fixed rather than remaining hidden. It also signals the school's commitment to equity in a concrete way.

Educational Institutions: Aligns with the institution's equity and inclusion mission. Schools and districts have equity goals; this control integrates those goals into AI usage. It also pre-empts potential crises – catching bias early prevents bigger issues like public scandals or loss of student trust. In terms of accountability, institutions can report that they are monitoring for and addressing bias, which could be important for school board reports, accreditation, or community trust. In some cases, it could shield the institution from liability – demonstrating due diligence in ensuring non-discrimination.

EdTech Developers: Although it might feel like scrutiny, this actually helps responsible developers by highlighting issues to fix, improving their product. For developers less aware of bias issues, it forces them to confront and resolve them or risk losing clients. It may also standardise reporting: if multiple schools demand bias audit results, developers might start providing "fairness reports" as part of their product package. In the bigger picture, it pushes the industry towards fairer AI. For the developer's reputation, if their product passes these school-run audits consistently, they can tout that as a quality indicator. Conversely, if a developer's product is repeatedly flagged for bias and they don't address it, they will lose trust and market share. So, there is incentive to cooperate with this control. Developers might also provide tools for educators to do some of this monitoring easily (like dashboards showing performance by subgroup) as a selling point.

10. Emotional Intelligence and Well-Being Safeguards

Definition

Monitor and support student emotional well-being in AI-mediated learning, with protocols for human intervention when needed. This control acknowledges that learning is an emotional process and that AI systems should not just track cognitive performance but also be attuned to signs of frustration, anxiety, or disengagement. It involves integrating tools or metrics for detecting student emotional states (e.g., if an AI tutor notices a student making repeated errors and taking longer pauses, it might infer frustration). However, it simultaneously emphasises privacy and non-intrusiveness – emotional monitoring should be done in a respectful, minimally invasive way (for instance, by analysing interaction patterns, not via creepy webcam eye-tracking without consent). Clear protocols must be established so that when an AI detects a possible emotional issue, it triggers a human response: e.g., notifying a teacher or counsellor, or suggesting the student take a break or seek help. The AI might also promote positive emotional engagement by incorporating encouragement, celebrating successes, and adjusting difficulty to avoid overwhelming the student. Essentially, this control ensures the AI contributes to, or at least does not harm, students' emotional and mental health during learning.

Challenges

Emotions are complex and vary widely among individuals, so detection is error-prone. An AI might misinterpret signals – one student's quietness is normal concentration, another's disengagement. False positives (flagging issues where there are none) could annoy or stigmatise students; false negatives (missing a student in distress) are even more concerning. Privacy concerns loom large: monitoring emotional state can feel invasive, and if any sensitive data (like mood or physiological data) is collected, it must be protected. Some parents might object to any form of emotional surveillance. Teachers might worry that this adds to their plate – getting pinged about a student's mood could be helpful, but if frequent or inaccurate, it could be overwhelming. Deciding when to intervene is also subtle; not every frustration needs escalation, sometimes struggling through is part of learning. Institutions need to delineate boundaries clearly: for example, will they attempt to detect serious issues like depression or self-harm through educational AI usage patterns? That veers into mental health territory with ethical implications. Also, aligning these safeguards with existing student support systems (counsellors, psychologists) requires coordination. Developers may have to incorporate affective computing techniques which are still evolving and not as mature as cognitive tutoring techniques. Ensuring these features work across diverse student populations (emotion expression can be culturally different) is tough. There's also a risk of overreach – we wouldn't want AI making psychological diagnoses; its role should be limited to flagging and supporting, which must be clearly defined.

Mitigation Strategies

- Use indirect indicators of emotional state that respect privacy. For example, track engagement metrics (response times, number of hints used, frequency of task switching). Sharp changes in these can indicate frustration or confusion. Avoid invasive methods like video emotion detection unless explicitly justified and consented (and even then, use cautiously given unreliability).

- Implement a multi-factor well-being assessment combining AI observation with student self-reports and teacher observation. For instance, an AI could occasionally ask the student "How are you feeling about this material?" and offer some simple choices (frustrated, okay, confident). Coupled with its own data, this self-report can improve accuracy. Teachers can also input their perceptions (maybe via a quick check-in form for each student). Using all three sources (AI metrics, student input, teacher input) will give a more robust picture and reduce over-reliance on the AI's guess.
- Recognise that each student has different behaviour patterns. The system could allow individual calibration. For example, a student with ADHD might always click quickly and appear disengaged by typical metrics, so their thresholds for flags might be adjusted. The framework could allow teachers to set different sensitivity levels or to override flags for certain students if they know those are that student's normal patterns. Essentially, avoid one-size-fits-all in determining what triggers an alert.
- Ensure the AI's interaction style promotes a positive emotional experience. Include encouraging messages for effort, not just correctness. For example, if a student is stuck, the AI might say, "This was a tough one – don't worry, many get it wrong. Let's try a different approach!" Small design elements like empathy in feedback and celebrating improvements can maintain morale and reduce frustration. By building an emotionally intelligent interface, the AI can sometimes defuse negativity before it escalates to needing teacher intervention.
- Develop a stepwise protocol for intervention when the AI flags an emotional concern. For instance: Step 1, AI gives the student a gentle prompt ("It seems you might be frustrated, would you like a hint or to take a short break?"). Step 2, if signs persist, AI notifies the teacher (or sends an alert to a dashboard). Step 3, teacher evaluates and decides if personal check-in or referral (e.g., to counsellor) is needed. Having this structured ensures human follow-up and that it's done consistently. It also sets boundaries – the AI tries minimal self-help first, then always involves a human, which is important for safety.

Implementation Guidance

Student mental health in digital learning has been a focus especially after experiences with remote learning. UNESCO (2023) warned of unprecedented risks to mental integrity if neurotechnology and AI are used without safeguards, underlining the importance of this control. Research by Luckin and Cukurova (2019) and others highlight that while AI can monitor certain signals, human intervention remains crucial for meaningful emotional support. That supports a design where the AI augments the teacher's awareness but doesn't take on the counsellor role fully. Al-Zahrani (2024) provides evidence of specific concerns about student well-being with AI and suggests detailed monitoring approaches and clear intervention thresholds. Implementation should probably start small: maybe pilot an AI feature that flags "high frustration" and see how accurate and useful teachers find it, then expand. Student and parent consent is a consideration: if any additional data (even something like "the student looked upset at 2pm on the app") is collected, it should be transparent. Possibly include an option for students to turn off emotion-related features if they feel uncomfortable (with caveats). Working closely with school counsellors or psychologists to frame what to monitor and how to respond is wise – they have expertise in child behaviour. These experts can help train teachers to interpret AI flags in context ("If the AI says student is disengaged, here are things to consider…"). Also, avoid pathologising normal behaviour: the idea is to catch when AI use causes or correlates with

genuine issues, not to label normal ups and downs as problems. On the tech side, simpler heuristics often work reasonably (like “3 wrong answers in a row quickly = frustration likely”) and are understandable to teachers, whereas complex emotion AI might be a black box. So, start with straightforward rules and only venture into advanced sentiment analysis if clearly beneficial. Data from the “Shape of the Future” indicated many school leaders are implementing protocols for regular well-being assessments in tandem with AI integration – often involving periodic surveys or check-ins separate from the AI. Our control suggests integrating some detection into the AI itself plus having those broader well-being checks. Importantly, maintain a strong privacy stance: any emotional data should be treated as sensitive and protected accordingly, and the purpose should strictly be to help the student. are vital to ensure fairness across different student populations, reinforcing that it’s not a one-time checkbox but a continuous dialogue. It’s also worth educating students and parents about what the school is doing to ensure fairness – this can build trust in the AI system. If an issue is found and corrected, transparently communicating that (with sensitivity to not erode confidence too much) can show the commitment to equity. Additionally, in some jurisdictions, algorithmic bias might have legal implications (anti-discrimination laws could apply if an AI systematically disadvantages a protected group), so this control also serves to keep the institution in compliance.

Relevance to Stakeholders

Students: Helps ensure that using AI for learning doesn't become a frustrating or demoralising experience. If a student is struggling alone with the software, this system will try to comfort or assist them and ultimately alert a human who can provide support. It can prevent feelings of helplessness or burnout by catching them early. Also, positive reinforcements from the AI can make learning more enjoyable and reduce anxiety. In sum, it safeguards their emotional well-being, which is as important as their academic success. Students are more likely to persevere and have confidence if they feel the "system" cares about how they feel, not just what score they got.

Teachers: Acts as an assistant in monitoring class well-being. A teacher can't always notice every quiet student or every frustrated face, especially in larger or online classes. The AI giving a nudge like "Maybe check on John, he's been inactive after multiple errors" can help teachers intervene at the right time. This complements teachers' own observations (they remain in control of deciding what to do). It can also provide data for mentors or counsellors about which students might need socio-emotional support. Ultimately, it helps teachers fulfil their role in nurturing the whole student, not just the academic part, by using tech as an extra set of eyes for the emotional climate.

Parents: Comforts parents to know that the school is mindful of the emotional impact of these technologies. Many parents worry about their child getting frustrated or overly stressed with new digital tools. Knowing there are features to catch and address that means their child is not left to struggle in silence. If a parent's child has particular emotional needs (say anxiety), they might appreciate that the system is tuned to flag if their child seems to be in distress so the teacher can respond. It shows the school values mental health and is integrating that concern even into tech usage.

Educational Institutions: Supports the institution's responsibility for student welfare. Schools are increasingly accountable for student well-being (surveys, mental health programs, etc.), and this ensures that edtech adoption doesn't run counter to those efforts. It reduces the risk of students having negative experiences with AI that could lead to disengagement from school or worse, emotional crises. If ever a concern arises (like "is AI hurting our students' well-being?"), the institution can point to these safeguards. It also fits into a broader push for Social-Emotional Learning (SEL) integration; AI can inadvertently erode SEL if isolating, but with this control, the AI can even become a tool that fosters resilience (by encouraging perseverance and seeking help appropriately).

EdTech Developers: Encourages building more empathetic AI systems, which could improve user satisfaction and outcomes. If students feel comfortable and supported by the AI, they'll likely use it more and get more benefit. It pushes developers into the realm of affective computing responsibly – not just detecting emotion for novelty, but to genuinely improve user experience. They may also differentiate their products with well-being features (some products now advertise that they incorporate mindfulness or encouragement). Developers do need to be careful with data privacy here; but if done right, being able to say "Our tutor has been proven to keep students engaged without frustration" is a selling point. Furthermore, working closely with educators on this can open new research and development avenues (like how to measure frustration through interaction patterns – a technical challenge with lots of ongoing research). Overall, it steers product development towards considering student mental health, which is a positive direction for the industry's social responsibility.

11. Organisational Accountability & Governance

Definition

Establish robust institutional oversight and clear lines of responsibility for AI systems used in education. This control entails creating governance frameworks – policies, committees, and processes – to ensure AI tools are deployed ethically and in compliance with legal requirements. It means assigning accountability at every stage of the AI lifecycle: from procurement and design to implementation and outcomes. In practice, schools or districts would designate roles (for example, an AI ethics officer or committee) to review AI integrations, monitor their performance, and address any issues. Likewise, EdTech providers must maintain corporate governance that aligns with these ethical standards, ensuring that when AI influences student learning or data, there is always a responsible human authority answerable for its behaviour and impacts. This proactive governance guarantees that AI decisions are transparent and that someone – whether a school administrator or a developer – can provide justification and take corrective action when the technology's outcomes are in question.

Challenges

Implementing organisational accountability faces technical, ethical, and institutional hurdles. AI systems can be complex "black boxes," making it hard for educators or administrators to understand how decisions are made, which complicates oversight. Without clear governance, if an AI unfairly marks student work or recommends a biased course of action, it may be unclear who is responsible for the error. Many educational institutions also lack established AI policies – recent research found roughly 40% of high schools surveyed had no AI-related guidelines at all (Ghimire & Edwards, 2024) – often due to limited expertise or resources. This policy vacuum means schools might adopt AI without an accountability structure, raising the risk of unchecked biases or privacy breaches. There can be ambiguity over roles: teachers might assume the district vetted a tool, while the district expects teachers to monitor classroom AI use. Such gaps allow issues to fall through the cracks. Moreover, defining how to hold AI accountable is still evolving; there isn't yet consensus on best practices, leading some organisations to take a cautious "wait-and-see" approach rather than pioneer strict governance (Hohma, 2023). Ethically, balancing innovation with control is tricky – too much bureaucracy might stifle beneficial AI experimentation, yet too little invites misuse. Developers face challenges as well: an EdTech company may be unsure how much transparency to provide (to satisfy school accountability demands) without exposing intellectual property. Finally, enforcing accountability can be arduous when AI is developed by third parties – a school might rely on a vendor's assurances and lack the capacity to independently audit the tool, creating dependence on the developer's own governance standards. All these factors make establishing clear accountability and governance an ongoing challenge in education settings.

Mitigation Strategies

- Set up a dedicated group (including school leaders, teachers, IT staff, and possibly parents or students) to oversee AI deployments. This committee can vet new AI tools for alignment with ethical guidelines and curriculum goals before they're adopted. Having a point person or officer for AI ethics ensures there is always someone with the mandate to monitor AI activities and champion accountability.

- Develop clear policies that outline how AI can be used and who is responsible for its outcomes. For example, a district might require an "AI Impact Assessment" before any system is used with students – a process to document the tool's purpose, data usage, and potential risks. Similarly, mandate periodic reviews or audits of AI performance (accuracy, bias, etc.) so that any drift or emergent issue is caught early. These reviews should involve educators and experts examining AI recommendations or grading patterns for fairness and appropriateness.

- Invest in training school staff and leadership on AI basics, ethical risks, and governance procedures. When teachers and administrators understand how the AI works and what could go wrong, they are better equipped to oversee it. For instance, train teachers to recognise when an AI may be making a faulty recommendation so they can override it, and train administrators on questions to ask vendors (e.g. about bias testing or data security). Building this internal capacity reduces over-reliance on vendors and creates a culture of shared responsibility.

- Incorporate strict accountability clauses into contracts with EdTech providers. Schools should insist on transparency from vendors – such as access to algorithmic audit results, documentation of how the AI was trained, and assurances of compliance with privacy laws. If feasible, use tools from providers that allow "explainable AI" features so educators can see why the AI suggested something. Establish channels for escalation: if a teacher or student reports a harmful AI behaviour, the vendor must have a support process to respond quickly (e.g. reviewing the incident, issuing a fix or guidance). By holding developers contractually accountable for ethical standards (bias mitigation, data protection, support), institutions extend their governance to the technology's source.

- Treat AI systems as part of an ongoing cycle of improvement. Implement dashboards or logs that track AI decisions and usage, which the governance committee can regularly review. Encourage teachers and students to provide feedback on AI tools – perhaps a quick report form if something seems off or if the AI isn't meeting needs. This bottom-up feedback is fed into governance meetings. If, for example, multiple teachers report the AI marking certain students unfairly, the committee can investigate and take action (adjust settings, work with the vendor, or even pull the tool from use until fixed). Having a defined feedback and remediation process means accountability is not one-off but sustained throughout the AI's deployment.

- Use emerging external frameworks to guide internal governance. Schools can look to national or international AI ethics guidelines (such as the EU's AI Act or IEEE standards) and adapt them to education. For instance, classify AI applications by risk level – a simple AI flashcard app vs. an AI making grading or disciplinary recommendations – and apply stricter oversight to higher-risk cases (similar to risk-based approaches in other industries). Ensure compliance with data protection regulations (like GDPR or FERPA) as a baseline, and go beyond minimum compliance by aiming for best practices (for example, adopting transparency and fairness principles from the OECD or UNESCO even if not legally required). By benchmarking against well-recognised standards, educational institutions can structure their governance in line with expert recommendations, and developers will likewise know what expectations they must meet.

Implementation Guidance

Educational institutions should take inspiration from both industry and policy developments to operationalise this control. A growing consensus holds that AI in high-stakes domains like education must be accompanied by formal governance. The European Union's draft AI Act explicitly classifies educational AI systems as "high-risk," meaning schools deploying AI for tutoring, grading or student analytics will likely be required to implement risk management, oversight, and documentation measures (UNESCO, 2024). Forward-thinking schools should start aligning with these practices now: for example, maintaining documentation of how an AI is used and decisions made, as one would do for other regulated processes. International bodies have also underscored accountability – the EU's expert group and UNESCO's AI ethics recommendations integrate accountability as a core principle for trustworthy AI. In practical terms, this means schools should institutionalise AI oversight rather than leave it ad hoc. One effective approach is borrowing the idea of multi-level governance: some universities (e.g. in the Big Ten Academic Alliance) convened cross-departmental committees to draft AI usage guidelines, involving IT, academic leadership, library science, and ethics experts (Wu et al., 2024). K-12 districts can emulate this by bringing together stakeholders – administrators, tech coordinators, teachers, perhaps board members – to collectively shape and enforce AI policies. Research suggests that such collaborative governance is valuable; a recent study noted that the lack of stakeholder involvement and clear guidelines in many high schools contributes to a "nascent governance stage" that leaves educators uncertain how to proceed (Ghimire & Edwards, 2024). Thus, engaging diverse voices (including teacher and parent representatives) when developing AI policies can build consensus and clarity, making implementation smoother.

Another key aspect is transparency and communication. Best practices from higher education and industry indicate that once policies are set, they must be clearly communicated to all levels of the organisation. Schools should issue plain-language guidelines for teachers on what the AI will and won't do, and instructions on their role in supervising it. Likewise, inform students (and parents) about how an AI tool is used in learning and what protections are in place – this can be done via student handouts or parent info sessions outlining the school's oversight measures. This openness not only builds trust but also reinforces accountability: everyone knows the rules and expectations, so it's easier to spot when something falls outside them. On the technical side, leveraging audit tools is advisable. For instance, if using an AI that grades essays, enable features that log its grading rationale or uncertainty levels, which the responsible teacher or committee member can review. In the finance sector, "model risk management" practices (regular audits, validation tests, bias checks) are standard – schools could adopt similar checklists for educational AI (e.g., test the AI on a variety of student groups to check for bias, verify that its recommendations match curricular goals, etc.). Partnering with external experts can help: a school might work with a university research centre to conduct an algorithmic audit or with a nonprofit to train the AI ethics committee on spotting issues. By setting up strong accountability mechanisms from the outset, educational organisations send a message to all stakeholders – and to AI vendors – that ethical, responsible AI use is a non-negotiable part of the educational mission.

Relevance to Stakeholders

Students: This control ultimately protects students' interests. When schools maintain accountability for AI, students are less likely to be subject to unchecked errors or biases in their learning tools. For example, if an algorithm unfairly flags a student's work or gives flawed feedback, a governance process ensures it gets corrected by a human before harming the student's grade or confidence. Students benefit from a trustworthy learning environment where AI is used as a tool to enhance their education rather than an opaque authority. In essence, Organisational Accountability means there's always a responsible adult watching out for students' rights and well-being when AI is involved. This builds student trust in the technology – they know there's recourse if something seems wrong – and reinforces that the purpose of AI is to help them learn, safely and fairly. It also sets a positive example for students about ethical technology use and the importance of responsibility in innovation.

Teachers: For educators, strong AI governance provides clarity and support. Teachers are often on the front lines using AI-driven apps or platforms; knowing that there is a clear policy and a support system gives them confidence. It means they have somewhere to turn if, say, an AI recommendation conflicts with their professional judgment or if they suspect the software is not working as intended. Rather than feeling that "administration dumped this AI on me without guidance," teachers become partners in implementation – often, governance frameworks invite teachers to give input and even be part of oversight committees. This inclusion elevates teacher agency: their observations can trigger reviews or improvements to the system. Moreover, accountability mechanisms ensure teachers are not unfairly blamed for AI missteps. If a homework grading AI makes an error, a transparent process will address it so the teacher isn't left solely responsible for the fallout. With clear governance, teachers also receive training and documentation, which helps them integrate AI tools more effectively into lessons.

Parents: Organisational accountability and governance are key to earning parent trust in educational AI. Parents are rightly concerned about who is "watching the watchers" when algorithms influence their children's learning or collect data. This control assures them that the school has put guardrails in place – there are policies, oversight committees, and contact points if problems arise. For instance, if a parent worries about an AI tutoring program's accuracy or bias, the school can explain the governance steps taken (such as vetting the content, regular audits, avenues for complaints). Knowing that the school leadership is actively monitoring AI tools – and ready to intervene if something goes wrong – reassures parents that these technologies won't operate unchecked. It also provides transparency: governance often includes reporting to the community, so parents might receive updates or be invited to forums about how AI is used responsibly at the school. In case of any incident (say a data breach or an AI misuse case), an accountable organisation will promptly inform parents and take responsibility, rather than leaving families in the dark. All of this builds confidence that AI in the classroom is being handled with the same duty of care that parents expect in all aspects of schooling.

Educational Institutions: For school and district leaders, this control is crucial for risk management and strategic oversight. By instituting governance, educational institutions ensure that the adoption of AI aligns with their educational mission and legal obligations. Governance also facilitates more effective decision-making: with committees and policies, leaders can make informed choices about which AI tools to allow, which to reject, and where to invest resources, based on systematic reviews rather than ad hoc decisions. This leads to more consistent, equitable technology use across classrooms. Additionally, demonstrating accountability can be advantageous when seeking funding or partnerships

– it signals to government bodies or grant organisations that the school is serious about responsible innovation. Ultimately, Organisational Accountability & Governance helps institutions harness AI's benefits (efficiency, personalisation, innovation) while safeguarding against its pitfalls, ensuring that the technology truly serves the school's pedagogical goals and values.

EdTech Developers: This control has significant implications for companies building AI for education. When schools demand accountability and have governance processes, developers are encouraged – and often required – to build products that meet higher ethical standards. This can be a challenge, but also an opportunity: developers who prioritise transparency (e.g. offering explainable AI features, detailed documentation), fairness (bias testing and mitigation), and privacy protections will find their products more readily accepted by institutions with strict governance. In effect, accountable schools push developers to "step up" their game, which can lead to better, safer products. Developers may need to engage more with stakeholders (running pilot programs, sharing data for third-party audits, responding to committee inquiries), but this collaboration can improve the AI's effectiveness and credibility. In the long run, adhering to strong governance requirements can become a selling point – EdTech firms can market that their AI has been vetted for bias or aligned with ethical guidelines, giving them an edge with conscientious buyers. Moreover, clear accountability distribution protects developers too: it clarifies what they are responsible for and what the institution handles.it steers product development towards considering student mental health, which is a positive direction for the industry's social responsibility.

12. Age-Appropriate & Safe Implementation

Definition

Ensure that AI tools and practices in education are tailored to students' developmental stages and uphold a safe, child-friendly learning environment. This control requires aligning AI usage with the age and maturity of learners, so that content, interactions, and capabilities are suitable and non-harmful. In essence, it's about designing and deploying AI with children's safety and well-being as a paramount concern. Practical examples include configuring AI tutoring systems with vocabulary and examples that match the reading level of a primary school student versus a teenager, or limiting certain AI functionalities for younger users (e.g. disabling open internet search or chat features in an elementary setting to avoid exposure to inappropriate material). It also involves robust content filtering and moderation – ensuring that any AI-generated content a student sees is free of violence, sexual content, hate speech, or other material not appropriate for their age. Safe implementation means the AI not only avoids harm but actively supports healthy development: for instance, encouraging positive social values, and not replacing developmental activities that children need (like play or face-to-face interaction). It covers privacy protections too, since handling data from minors must be done with extra care.

Challenges

Adapting AI to be age-appropriate and safe for children presents multiple challenges. One major difficulty is content control: AI models, especially generative ones, can produce unpredictable outputs. Without rigorous filters, a well-meaning educational chatbot could inadvertently show a 10-year-old content meant for adults or misinformation that the child isn't equipped to vet. Ensuring 100% safe content is technically challenging – filters can sometimes be too lax (letting bad content slip through) or too strict (blocking legitimate educational material). Developmentally appropriate interaction is another hurdle: young children think and communicate very differently from teenagers. An AI that works well for a high schooler (e.g. a complex reasoning assistant) might confuse or even frighten a younger child with long-winded or overly technical responses. Tuning the AI's language and approach for different ages requires careful design and often multiple versions of a model, which is resource-intensive. There's also the risk of overtrust and misunderstanding. Children, especially younger ones, might not grasp that AI has limitations – they could take an AI's answers as always correct or even form emotional attachments to a friendly sounding AI character. Studies have found that children can treat AI chatbots or robots as if they were human friends, even confiding personal feelings or secrets to them Kurian, N. (2024). This trust can be dangerous if the AI gives poor advice or if the child divulges private information. It places a burden on the system to handle such situations appropriately – something current AIs are not fully capable of, since they cannot truly care or intervene like an adult would. From an ethical standpoint, privacy is a big concern: AI systems often collect data to function (e.g. learning progress, personal preferences), and doing so with minors triggers legal requirements (like parental consent under laws such as COPPA) and moral obligations to guard that data tightly. Schools and developers might struggle to navigate these regulations and to build systems that use minimal data to achieve their goals, which is the safest route.

Institutionally, implementing age-based distinctions can be complex. A school might have to maintain different tool settings for different grade levels, which is technically and logistically demanding (e.g. ensuring a 4th grader using an app gets the child-safe mode,

whereas a 10th grader can utilise more open features). If the AI doesn't automatically adapt to age, there's reliance on busy teachers or IT staff to configure it properly for each class – mistakes could lead to a student getting the wrong experience. Additionally, one-size-fits-all solutions are hard: even within the same age. Lastly, developers often face a lack of child-specific data or research to guide design. AI models are typically trained on general data (largely from adult interactions), so they may not naturally handle child inputs (like the whimsical, imprecise language kids use). Adapting models to kids can require additional training data that is hard to obtain due to privacy or simply because kids communicate differently in supervised research settings. All these challenges – technical, regulatory, and practical – make it a non-trivial task to implement AI that is both safe and suitably tailored for each age group in education.

Mitigation Strategies

- Implement clear age-based tiers for AI usage. For example, have a "junior" mode versus "senior" mode in an educational app. Younger students might get a very restricted version of the AI – with a limited set of functions and heavily pre-curated content – while older students have more freedom. This could involve requiring a teacher or parent to authenticate or unlock advanced features for a student above a certain age. By gating content and capabilities, you prevent children from straying into functionalities that aren't meant for them. Many general platforms already do this; schools and EdTech providers can mirror that concept for AI educational tools.

- Use multiple layers of filters to catch inappropriate content. This includes keyword-based filters (to block profanity, sexual terms, violence references), AI moderation models that detect hate speech or self-harm content, and human review for any predefined content library. For generative AI that produces answers or stories, integrate a safety module that reviews the output before it reaches the student. If the content is deemed unfit or even just ambiguous, the system can either refuse to answer or flag it for teacher review. It's also wise to maintain an updated blacklist/whitelist – for instance, explicitly ban any websites or sources known to be unreliable or non-child-friendly from the AI's web access. On the flip side, ensure the AI actively includes diverse, positive content appropriate for children (e.g., examples from children's literature, age-appropriate cultural references) to keep the experience engaging without venturing into unsafe territory. Continually update these filters based on real-world use: if a new form of slang or meme with bad meaning emerges among kids, the moderation system should learn to catch it.

- Tailor the AI's interaction style to the cognitive level of the user. This might mean designing separate conversational datasets: one full of simplified language, cheerful encouragement, and step-by-step guidance for young learners, and another with more complex, nuanced language for older students. The AI should adjust things like sentence length, vocabulary, and the complexity of concepts based on either the student's age or demonstrated ability. For young children, the AI might incorporate more storytelling or gamified elements (since play is crucial at that stage), whereas for teenagers, it can be more direct and allow deeper critical discussion. Testing the AI with target age groups is key – gather feedback from actual students in different grades to see if they find it understandable and comfortable.

- Give parents and educators control and insight into the AI's use. For instance, allow parents to opt their child out of certain AI features or to receive summaries of what their child is doing with the AI (similar to how some internet filters send weekly

reports). In a classroom setting, a teacher dashboard for an AI app can show, in real-time or via logs, what kinds of questions students are asking and the AI's responses. This transparency means an adult can audit or spot-check the interactions to ensure they remain appropriate. Provide an easy way for parents or teachers to flag any content or AI behaviour they find concerning – a simple "Report this response" button that triggers a review by the developer's team and notifies the school. By involving parents, you not only reassure them but also leverage their eyes and ears to maintain safety. Clear communication is part of this strategy: inform parents at the start of the school year, "We will be using AI tool X in the classroom, which has been configured for children of this age. Here's what it does, and here are the safety measures in place." When parents know the school is mindful of age-appropriateness, they are more likely to consent to and support AI initiatives.

- Minimise data collection and enable strong privacy by design. For any AI system used by under-18 students, ensure it collects only what is pedagogically necessary. Avoid using personal identifiers unless required; use anonymised or local profiles if possible (e.g., the AI can function with a nickname or student ID and doesn't need full personal details). If the AI uses student data to personalise learning, keep that data encrypted and inaccessible to any external parties. Obtain explicit parental consent for data usage in compliance with laws – and even when legally not mandated (for older minors), consider it a best practice to be transparent and get buy-in. Set retention limits so the system doesn't keep a child's data indefinitely; for example, delete or aggregate data after a school year or when no longer needed. Moreover, build in protections against the AI eliciting personal data from a child. The AI should be programmed not to ask for personal information like address, full name, or contact info – and if a student volunteers such info, the AI could respond with a gentle warning ("I don't need to know that to help you, let's keep our conversation about school subjects!"). This aligns with standard child-safety rules and prevents exploitation.

- Even the best-designed AI requires informed human guidance, so implement training for both students and educators on safe AI practices. Teach students, in an age-appropriate way, about what AI is and isn't – for example, a simple lesson that "the classroom helper app is just a computer program, it might make mistakes, and you should always feel okay telling a teacher if something it says makes you uncomfortable." Empowering students with some digital literacy can prevent blind overtrust. Simultaneously, train teachers on how to integrate the AI safely: for young kids, maybe the teacher always initiates or supervises AI sessions; for older ones, set rules like "don't use the AI to get answers you wouldn't normally be allowed to look up." By establishing usage norms (much like internet safety rules), schools create a culture where AI is a monitored tool, not a free-for-all. Additionally, incorporate check-ins: if using AI regularly, teachers might have weekly brief discussions with the class about their experiences – "Did the AI say anything weird or confusing to anyone this week?" – to surface any issues early.

Implementation Guidance

Implementing age-appropriate and safe AI in education should build on emerging best practices from child psychology, educational research, and child-rights frameworks. A foundational step is to refer to established guidelines like UNICEF's Policy Guidance on AI for Children (2021), which underscores that AI systems should support children's development and rights, prioritise safety, and use age-appropriate language and transparency when interacting with kids. In concrete terms, this means developers and

educators should collaborate to make AI explain its purpose and responses in ways a child can understand – for instance, an AI tutor might have a friendly avatar that can say, "I'm a computer program here to help you learn. If I say something that seems wrong, you can ask your teacher!" Such explanations, recommended by child-focused AI guidelines, help set correct expectations. Organisations like UNESCO have also weighed in: UNESCO's 2023 global guidance on AI in education emphasises strict safeguards for minors, even suggesting a minimum age of 13 for using certain AI tools in the classroom and calling for dedicated teacher training on these tools' ethical use. This aligns with many jurisdictions' approach (for example, social media and online services often restrict under-13 users due to maturity and privacy concerns). Schools implementing this control should consider these age limits and ensure younger students only access AI under close supervision or in very controlled formats. Notably, where younger children do use AI, it should be with full consent and knowledge of parents, and ideally using platforms specifically designed for children (rather than generic AI apps).

From a design perspective, leveraging the concept of "Child-centred Design" is key. This approach, advocated by experts in human-computer interaction, involves including children and educators in the development loop of AI products. For instance, before rolling out an AI reading assistant to all second-graders, pilot it in one class and observe how the children interact with it – their confusion or delight will tell designers what to tweak. Many edtech companies are starting to form advisory panels with teachers and child psychologists to review AI content for appropriateness. Schools can encourage this by favouring products that can demonstrate they followed such processes. In some cases, districts themselves partner with researchers; for example, a school might work with a university education department to evaluate whether an AI's output aligns with developmental benchmarks (does it foster critical thinking in a 15-year-old appropriately? Is it boosting vocabulary in a 7-year-old without introducing concepts that are too advanced?). These partnerships can generate case studies that benefit the wider community.

There are also technical standards and certifications emerging for kid-safe tech. The "Age Appropriate Design Code" (also known as the Children's Code in the UK) established 15 principles for online services to protect children's data and well-being. Educational AI platforms would do well to adhere to such principles globally, even if not legally required in their region, as a mark of safety-by-design. Schools implementing AI should ask vendors about compliance with such standards: for example, does the platform have a high default privacy setting for students? Can the AI's interface adapt to different age ranges? A practical tip is to maintain documentation of how each AI tool was configured for safety. If an inspector or concerned parent asks, the school can show, for instance, "Tool X is used in middle school with SafeMode on, profanity filter active, web access off," demonstrating due diligence.

In terms of everyday practice, it helps to draw parallels with existing measures schools take. Think of how field trips require permission slips and age-suitable planning – similarly, using a new AI app might require a "permission and info sheet" to parents and a pilot with a small group of students. Consider also the lessons learned from internet usage in schools: many schools implemented web filters and taught digital citizenship to students when the internet became ubiquitous. AI is analogous – content filters (as detailed in mitigation) and AI literacy education should go hand in hand. In fact, experts suggest incorporating AI literacy into curriculum from an early age, so students learn early that not everything an AI says is true and that they should question and verify information (Munzer, 2024). By fostering a bit of healthy scepticism and critical thinking, students become safer users.

A recent Harvard study on children and AI noted that kids need guidance to interpret AI's behaviour (since AI might not follow normal human social rules or could lack empathy) (Xu, 2024). Teachers can use guided discussions or role-playing exercises (like "pretend the AI said something mean – what should you do?") to reinforce that the AI isn't a person and that certain responses are inappropriate and should be reported.

Finally, continuous improvement is crucial. Safe implementation is not a one-time setup but an ongoing process. Schools should schedule periodic evaluations of the AI tools in use – for example, at the end of each semester, review if the content filtering has been effective and update it if any issues were reported. Keep an eye on updates from AI providers: if a new version of the AI model is released, re-test it for age-appropriateness because changes in the model could introduce new types of responses. Engage with the student voice as well; perhaps have a student council or focus group (especially for older students) give feedback on how they feel using the AI tools. Their perspective can be invaluable – they might point out, for instance, that the tone of an AI mentor app feels too childish for 11th graders, which could be adjusted to maintain engagement. On the flip side, younger kids might say they love a maths AI that uses cartoon avatars, which validates that the child-friendly design is working. By treating age-appropriate AI use as an evolving practice, and staying informed on research and guidelines, educators can ensure that as AI tools grow in capability, they do so under the protective umbrella of policies and design strategies that put children's safety first.

Relevance to Stakeholders

Students: Enforcing age-appropriate and safe AI means students get the maximum benefit from AI tools with minimum risk. For younger children, it creates a learning space that is engaging and supportive without exposing them to scary or confusing content – the AI becomes like a friendly tutor who speaks their language. This helps students learn more comfortably; they're not bewildered by instructions that are too advanced or traumatised by something inappropriate popping up. It also subtly teaches them in a developmentally fitting way – for instance, a child-safe AI might encourage a 7-year-old with simple praise ("Great job, you solved the puzzle!") which is exactly the kind of positive reinforcement that age group needs, whereas a teen-focused AI might give a 17-year-old more analytical feedback ("Check your second step, there might be an error in your algebra") which respects their growing autonomy. By having these safeguards, students are less likely to encounter harmful situations, like cyberbullying from a misuse of AI or dependency on an AI that does all their thinking.

Teachers: For educators, having AI that is age-appropriate and safe is essential to integrate it confidently into teaching. It reduces the "fear factor" that something might go wrong. A teacher can allow a classroom of 3rd graders to use a reading assistance AI without hovering in panic, because they know the tool has been vetted and locked down to a safe mode. This frees teachers to focus on facilitating learning rather than constantly policing the AI. When issues do arise (say a student gets an odd response), teachers, being in the loop through dashboards or reports, can swiftly intervene and use it as a teachable moment. Age-appropriate design also means the AI aligns with curriculum standards for that grade, which helps teachers meet their learning objectives.

Parents: This control is perhaps most visibly reassuring to parents. Parents often have deep concerns about digital tools in the classroom – from screen time to exposure to harmful content or strangers online. By emphasising age-appropriate and safe implementation, schools send a signal to parents that "we prioritise your child's safety above all when using AI." For instance, if a parent hears that the school is rolling out an

AI-powered maths tutor, their first question might be: is this safe for my child? With this control in place, the school can answer yes, explaining the filters, the limited features, and the supervision in place. Knowing that there are strict safeguards (like no collection of personal data, no chatting with unknown entities, and content tailored for kids) puts parents more at ease.

Educational Institutions: For schools and districts, age-appropriate and safe implementation of AI is critical to fulfil their duty of care and maintain their educational integrity. Legally and ethically, institutions are expected to protect minors in their charge – this means if they introduce AI and something harms a student, they could face serious consequences. It aligns with initiatives like digital citizenship and SEL (Social-Emotional Learning) programs schools often run: integrating AI safely complements these by ensuring technology doesn't undermine those efforts (for example, an AI that encourages collaboration and kindness rather than exposing kids to toxic online behaviour). There's also a practical operational benefit: fewer disruptions. If AI is rolled out without safety in mind, schools may find themselves constantly firefighting – dealing with upset parents, retracting tools, or even handling trauma if a student was seriously affected by something. Safe implementation avoids these derailments, allowing the institution to focus on the constructive use of AI.

EdTech Developers: For companies building AI tools for education, focusing on age-appropriate and safe design is not just about being socially responsible – it's increasingly a market expectation. Schools and parents are more likely to adopt and stick with products that demonstrate a commitment to child safety. Developers who invest in these features (like curated content libraries for different ages, comprehensive profanity filters, or adjustable reading levels) will stand out in a crowded edtech market. There is a cost to implementing this control on the development side – it may require more content moderation staff, refined algorithms, and consultations with experts in children's media. However, it also opens opportunities for innovation: some companies might develop proprietary child-friendly datasets or novel ways to simplify AI explanations, which then become part of their intellectual property advantage.

Conclusion

With this initiative we hoped to offer a comprehensive starting point for responsibly integrating AI into education. By focusing on principles such as user agency, cultural sensitivity, bias mitigation, and transparent communication, the twelve Ethical Controls seek to ensure that the benefits of AI—greater personalisation, efficiency, and access—are realised without undermining core educational values or learners' well-being. Taken together, these controls should provide a structured yet flexible set of guidelines that schools, educators, policymakers, and developers can adopt and adapt to their own unique contexts.

However, this framework must be understood as a "living" resource rather than a static rulebook. AI is evolving at a rapid pace, and the ways it affects learning will inevitably shift over time. New technologies will emerge, student populations will change, and the very definition of "responsible AI" will continue to evolve as societal and ethical expectations grow more sophisticated. Consequently, the guidelines outlined should be revisited regularly to accommodate new insights, address unforeseen risks, and capture promising opportunities.

Equally important is active engagement with all stakeholders—students, teachers, parents, educational leaders, and developers—so that the framework is continuously refined in practice. Feedback loops, pilot programs, and impact evaluations will be essential to updating the controls in real-world contexts. By maintaining a shared commitment to iterative improvement, the educational community can ensure that AI's growth remains aligned with human-centered values and genuine learning outcomes.

In this sense, implementation of the framework marks a milestone rather than a final endpoint. It creates a shared language to foster collaboration, offering a practical roadmap to guide responsible innovation. As organisations put the twelve controls into action, their experiences and findings will in turn shape the next generation of ethical guidelines, ultimately helping AI reach its highest potential for positive transformation in education.

Appendix

Case Studies

The University of Sydney Model

The University of Sydney's implementation of AI controls provides a framework built upon four foundational principles: establishing rules, providing equitable access, building familiarity, and fostering trust. Their approach is particularly noteworthy for its systematic development of governance structures and practical implementation guidelines that address both immediate and long-term challenges of AI integration in education.

Central to their implementation was the establishment of a robust governance structure. The University created a Generative AI Steering Committee, co-chaired by the Deputy Vice-Chancellor (Education) and Deputy Vice-Chancellor (Research), demonstrating the institution's commitment to integrating AI considerations across both educational and research domains. This committee meets monthly and reports directly to the University Executive and Senate, ensuring consistent high-level oversight of AI implementation. Supporting this primary committee is a coordinating group that includes representatives from research, education, ICT, library services, and student administration, providing comprehensive stakeholder representation in the decision-making process.

The University developed a set of clear "guardrails" to guide AI implementation, focusing particularly on data protection, privacy, and intellectual property concerns. These guardrails serve as practical guidelines for both staff and students, encouraging experimentation with AI while ensuring appropriate protections for sensitive information. The framework emphasises the importance of transparent acknowledgment of AI use and includes specific protocols for handling different types of institutional data.

Perhaps most innovative is the University's development of a "two-lane" approach to assessment, which directly addresses the challenges of maintaining academic integrity in an AI-enabled environment. Lane 1 consists of secured, in-person, supervised assessments that serve to verify student learning outcomes, while Lane 2 encompasses unsecured assessments that focus on the learning process itself. This bifurcated approach is supported by an "AI x Assessment Menu" that provides faculty with specific guidance on incorporating AI into their teaching and assessment practices.

The University's framework places significant emphasis on building familiarity with AI technologies among all stakeholders. This includes mandatory training modules for staff, targeted transition activities for first-year students, and regular workshops and professional development opportunities. The Education Portfolio, in collaboration with the Library, has developed comprehensive AI literacy programs and resources, including a 20-minute introductory activity embedded in first-year transition units.

The Singapore Education System

Singapore's approach to AI implementation demonstrates how controls can be effectively integrated into a national education framework. Their system is particularly notable for its emphasis on learning analytics and practical applications within a broader national technology strategy.

Singapore's structured oversight begins with integration into their Smart Nation Initiative, providing a comprehensive framework for AI implementation across the education sector. This integration extends to the development of the Singapore Student Learning Space (SLS), a national-level learning management system that incorporates AI-enabled adaptive learning capabilities while maintaining appropriate controls and oversight.

The Singapore model places particular emphasis on learning analytics and data-driven decision-making. Their framework includes the development of sophisticated descriptive analytics tools and visualisation systems, exemplified by projects such as AppleTree and CoVAA, which provide real-time learning analytics while maintaining appropriate privacy and security controls. These systems demonstrate how AI can be used to enhance student engagement and learning outcomes while operating within clear ethical boundaries.

A distinctive feature of Singapore's approach is its emphasis on cultural sensitivity and ethical considerations in AI implementation. Their framework includes regular cultural audits of AI systems and emphasises stakeholder engagement throughout the development process. This attention to cultural and ethical considerations has resulted in the development of specific controls for bias detection and mitigation, ensuring that AI implementations serve the needs of Singapore's diverse student population.

The AMIA Healthcare Model: Transferable Controls for Education

The American Medical Informatics Association's (AMIA) framework for AI governance, while developed for healthcare settings, offers insights for educational control development. Their approach is particularly notable for its structured emphasis on ethical principles and practical governance mechanisms that can be effectively adapted for educational contexts.

Central to the AMIA framework is a set of ethical principles derived from traditional medical ethics but applicable to AI implementation. The framework emphasises four fundamental principles: beneficence (explicitly designing AI to be helpful), nonmaleficence (preventing harm), autonomy (protecting individual choice and agency), and justice (ensuring equitable access and representation). These principles provide a robust ethical foundation that can be readily adapted to educational contexts.

The framework's approach to technical implementation is particularly noteworthy for its emphasis on trustworthiness, which is divided into organisational and technical principles. On the organisational level, the framework emphasises benevolence (developing AI for positive purposes), transparency (clear communication about AI capabilities and limitations), and accountability (active oversight and risk management). These organisational controls ensure that AI implementation aligns with institutional values while maintaining appropriate oversight.

Technical principles within the AMIA framework include several elements crucial for educational adaptation. The requirement for explainability ensures that AI systems can be described in context-appropriate language, making their scope and limitations understandable to all stakeholders. Interpretability requirements ensure that systems can provide plausible reasoning for their decisions or advice in accessible language. The framework also emphasises the importance of fairness, requiring systems to be free of bias and non-discriminatory, with regular auditing processes to ensure compliance.

A particularly valuable aspect of the AMIA framework is its approach to lifecycle management of AI systems. The framework outlines specific controls for each stage of

AI implementation, from inception through deployment to eventual decommissioning. This includes requirements for rigorous testing during development, comprehensive documentation of system capabilities and limitations, and clear protocols for ongoing maintenance and eventual system retirement.

The framework places significant emphasis on professional development and support, requiring comprehensive training programs for all stakeholders. This includes education about AI capabilities and limitations, ethical considerations, and practical guidelines for appropriate use. The framework also emphasises the importance of ongoing support rather than just initial training, ensuring that users maintain currency with emerging capabilities and challenges.

Data privacy and security considerations are particularly robust in the AMIA framework, with specific controls for handling sensitive information. The framework requires clear protocols for data protection, including requirements for audit trails, security measures, and privacy safeguards. These controls can be readily adapted for educational contexts, particularly regarding student data protection and privacy considerations.

The framework's approach to vulnerable populations provides valuable guidance for educational implementations. It emphasises the need for increased scrutiny and additional safeguards when AI systems interact with vulnerable groups, a consideration particularly relevant in educational contexts where students may be minors or otherwise vulnerable.

The AMIA framework also provides valuable insights into implementation challenges and resource requirements. It acknowledges the significant investment required in technical infrastructure, personnel resources, and time, while providing frameworks for scaling implementation based on available resources while maintaining essential ethical safeguards.

A final notable aspect of the AMIA framework is its emphasis on continuous improvement and adaptation. The framework requires regular assessment of AI systems' impact, ongoing monitoring of emerging risks and challenges, and periodic updates to controls and guidelines. This approach ensures that governance remains responsive to evolving technology while maintaining consistent ethical standards.

British Columbia K-12 Framework: Considerations for AI Implementation

The British Columbia Ministry of Education and Child Care's framework for AI implementation in K-12 schools provides a comprehensive approach specifically designed for primary and secondary education contexts. The framework is particularly notable for its role-specific guidance and detailed consideration of practical implementation challenges in school settings.

The framework establishes a multi-tiered approach to AI governance, recognising the distinct needs and responsibilities of different stakeholder groups within the education system. It provides specific guidance for school boards, district leaders, school leaders, and teachers, ensuring that each group understands their role in responsible AI integration. This layered approach allows for consistent implementation while accommodating local needs and priorities.

A key strength of the framework is its emphasis on responsible integration of AI tools within existing educational structures. The framework explicitly recognises that education is inherently relational, positioning AI as a complement to human processes rather than a replacement. This foundational principle shapes all aspects of the framework's

implementation guidance, ensuring that human connections remain central to the learning process.

Core Implementation Categories

The framework identifies seven distinct categories for consideration when implementing AI in schools:

- Ethical Uses
- Needs and Impacts
- Accessibility and Usability
- Integration and Compatibility
- Data Security and Privacy
- Teaching and Learning
- Inclusive Learning

Each category is approached with detailed consideration of practical implementation challenges specific to K-12 environments. For instance, the framework emphasises the importance of transparent evaluation processes during selection, implementation, and decommissioning stages of AI tools, recognising the full lifecycle of technology adoption in schools.

Privacy and Data Protection

The framework places particular emphasis on privacy legislation and protection of student information. It requires schools to conduct Privacy Impact Assessments (PIAs) when implementing AI tools, ensuring compliance with Freedom of Information and Protection of Privacy Act (FOIPPA) requirements. This systematic approach to privacy protection includes specific guidance on data collection, use, and protection of personal information.

Professional Development and Support

A significant focus of the framework is building capacity among educators and staff. It emphasises the importance of thoughtful assessment of professional learning opportunities within schools and districts before selecting and implementing AI tools. The framework specifically calls for the development of AI literacy among all district employees, recognising that successful implementation requires broad-based understanding of AI capabilities and limitations.

Cultural and Equity Considerations

The framework places strong emphasis on cultural sensitivity and awareness of diverse perspectives, including Indigenous ways of knowing. It requires consideration of how various individual, social, and environmental differences can influence students' ability to access and benefit from AI tools. This includes specific attention to socioeconomic differences, disabilities and diverse abilities, and barriers associated with colonisation.

Assessment and Evaluation

The framework emphasises the importance of continuous assessment of AI implementation impact. It requires regular evaluation of both efficiency and effectiveness,

including consideration of whether AI tools actually improve student learning outcomes. The framework specifically calls for monitoring of AI tool usage and prompt discontinuation if tools fail to meet specific needs of schools, districts, or classrooms.

Accessibility and Inclusivity

A distinctive feature of the framework is its comprehensive approach to accessibility and inclusivity. It requires consideration of how AI tools can accommodate diverse learning styles and individual needs, emphasising the importance of equitable access. The framework specifically addresses the need to bridge any gaps in access between students, recognising that additional support and services may be required to reduce barriers for certain students.

References

Anderson, J. (2024, October 2). *The impact of AI on children's development*. Harvard Graduate School of Education. https://www.gse.harvard.edu/ideas/edcast/24/10/impact-ai-childrens-development

Baker, R. S., & Hawn, A. (2021). *Algorithmic bias in education. International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Le Quy, T., & Zhang, W. (2024). *FairAIED: Navigating fairness, bias, and ethics in educational AI applications*. arXiv. https://arxiv.org/abs/2407.18745

Eaton, S. E. (2023). *Academic integrity and artificial intelligence: An ethical framework for educators and researchers*. International Journal for Educational Integrity, 19(1), Article 17. https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00144-1

European Parliament & Council of the European Union. (2024, June 13). *Annex III: High-risk AI systems referred to in Article 6(2). In Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/annex/3/

Foltýnek, T., Dlabolová, D., Glendinning, I., Lancaster, T., & Linkeschová, D. (2023). *ENAI Recommendations on the ethical use of Artificial Intelligence in Education*. European Network for Academic Integrity. https://www.researchgate.net/publication/370455881

Ghimire, A., & Edwards, J. (2024). *From guidelines to governance: A study of AI policies in education*. arXiv. https://arxiv.org/abs/2403.15601

Holstein, K., McLaren, B. M., & Aleven, V. (2020). Designing for human–AI complementarity in K-12 education. *Artificial Intelligence*, *50*(6), 1–15. https://doi.org/10.1002/aaai.12058

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, *50*(6), 2824–2838. https://doi.org/10.1111/bjet.12861

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, *12*(8), Article 153. https://www.mdpi.com/2073-431X/12/8/153

Munzer, T. (2024, February 22). *What is age-appropriate use of AI? 4 developmental stages to know about*. University of Michigan, Department of Pediatrics. https://medicine.umich.edu/dept/pediatrics/news/archive/202402/what-age-appropriate-use-ai-4-developmental-stages-know-about

Pane, J. F., Griffin, B. A., McCaffrey, D. F., Karam, R. T., Daugherty, L., & Phillips, A. (2013). *Does an algebra course with tutoring software improve student learning?* RAND Corporation. https://www.rand.org/pubs/research_briefs/RB9746.html

Sullivan, M. (2023, September 8). *NIST AI Risk Management Framework: A comprehensive overview*. Transcend. https://transcend.io/blog/nist-ai-risk-management-framework

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

UNESCO. (2023, November 6). *Use of AI in education: Deciding on the future we want*. https://www.unesco.org/en/articles/use-ai-education-deciding-future-we-want

UNICEF Switzerland and Liechtenstein. (2023). *Policy guidance on AI for children*. https://www.unicef.ch/sites/default/files/2023-06/AI%20Statement%20EN_050623.pdf

Wu, C., Zhang, H., & Carroll, J. M. (2024). AI governance in higher education: Case studies of guidance at Big Ten universities. *Future Internet*, *16*(10), Article 354. https://www.mdpi.com/1999-5903/16/10/354

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0