THE Safety Chic

MAKING CHILD-CENTRED TECHNOLOGIES SAFER:

AI, DATA ETHICS, AND INNOVATION STANDARDS

July 2025

THE Safety Chic

Website: www.thesafetychic.com
Contact: info@thesafetychic.com

Contributors and Reviewers

Amarachi Iheagwam, Mercy John, Joshua Mze, Chidimma Obidiegwu, Ugochi Obidiegwu, Olufunmilayo Lasisi, and Tolu Sogbesan

Acknowledgements

As rapid progress occurs in digital technology development and AI use in almost all products, I am concerned about the potential impact on safety especially for children. This concern led to further research and this white paper. I would like to thank all staff, interns, and partners at The Safety Chic who supported the delivery of this paper. Thank you for your generous expertise and time throughout the writing process. I also extend our sincere gratitude to all the individuals and organisations whose insights, research, and feedback contributed to the development of this white paper. I hope these insights help in creating safer digital innovations for children.

Ugochi Obidiegwu
Founder, The Safety Chic

TABLE OF CONTENT

LIST OF ACRONYMS

AI	Artificial Intelligence
CAI	Conversational Artificial Intelligence
CBT	Cognitive-Behavioural Therapy
COPPA	Children's Online Privacy Protection Act
DFC	Digital Future for Children
ECE	Early Childhood Education
EU	European Union
GAI	Generative Artificial Intelligence
GDPR-K	General Data Protection Regulation for Kids
ICO	Information Commissioner's Office
IEEE	Institute of Electrical and Electronics Engineers
LLM	Large Language Models
NLP	Natural Language Processing
UI/UX	User Interface/User Experience
UNICEF	United Nations Children's Fund
XAI	Explainable AI

EXECUTIVE SUMMARY

The accelerating integration of Artificial Intelligence (AI) into technologies designed for children, ranging from interactive educational platforms and personalised learning apps to smart toys and digital companions, presents a transformative paradigm for child development and engagement. These innovations offer unprecedented opportunities for tailored learning experiences, enhanced cognitive development, and dynamic social interactions. However, this rapid technological evolution also introduces a complex array of profound ethical challenges, particularly concerning the inherent vulnerabilities of children. Their developing cognitive abilities, limited understanding of complex data processes, and often trusting nature make them uniquely susceptible to risks associated with extensive data collection, algorithmic biases, and subtle forms of digital manipulation. This white paper critically examines the dual nature of AI in child-centric technologies, meticulously exploring its vast potential alongside the significant ethical concerns that demand urgent attention and proactive mitigation.

The core ethical dilemmas explored include the pervasive issues of data privacy and security, where children's personal information is often collected without adequate transparency or informed consent, posing substantial risks of misuse, surveillance, and long-term implications for their digital footprint. Equally critical is the concern of algorithmic bias. When AI systems are trained on unrepresentative or flawed datasets, they can inadvertently perpetuate and amplify societal stereotypes, leading to unfair or discriminatory outcomes that may limit opportunities or reinforce harmful narratives for certain groups of children. Furthermore, this paper addresses the nuanced threat of manipulation and persuasion, as AI's ability to analyse and adapt to individual behavioural patterns could be exploited to encourage excessive screen time, influence purchasing decisions, or subtly shape developmental pathways without explicit awareness from children or their guardians.

Addressing these ethical quandaries necessitates a comprehensive and proactive approach rooted in robust ethical frameworks and vigilant governance. This white paper advocates for several crucial strategies: prioritising transparent data practices that are easily understandable by both children and their guardians, ensuring privacy by design and security by default in all child-centred AI technologies, and developing child-friendly AI design principles that place the child's best interests and developmental stage at the forefront. It underscores the imperative for comprehensive legal and regulatory frameworks that are agile enough to keep pace with technological advancements, alongside fostering multi-stakeholder collaboration among parents, educators, technology developers, policymakers, and civil society. Successful initiatives from UNICEF and various national efforts towards digital literacy and ethical AI development serve as beacons, illustrating effective models for fostering a safe, equitable, and empowering digital future where AI truly serves the holistic well-being of children.

ABSTRACT

The rapid integration of Artificial Intelligence (AI) into child-focused technologies ranging from educational platforms and smart toys to digital companions, presents both transformative opportunities and urgent ethical challenges. While AI promises to enhance learning, development, and engagement, it also exposes children to risks such as data exploitation, algorithmic bias, and digital manipulation, given their unique vulnerabilities. This white paper critically examines the ethical landscape of AI in child-centred technologies, highlighting concerns around data privacy, biased algorithms, and persuasive design. It advocates for a rights-based, proactive approach that includes transparent data practices, privacy-by-design, child-sensitive AI frameworks, and agile regulatory oversight. The paper emphasises the importance of collaboration among parents, educators, developers, and policymakers. It also calls for concerted action to ensure that AI technologies safeguard, rather than compromise the well-being and agency of children in the digital age.

INTRODUCTION

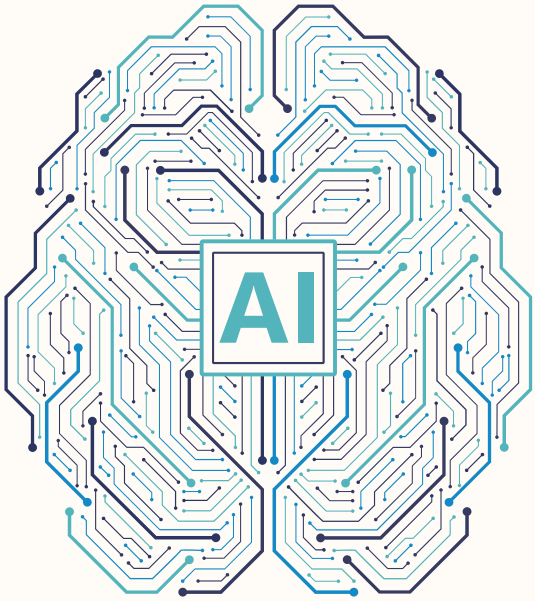
As artificial intelligence (AI) becomes increasingly integral to the tools children use for learning, play, and communication, the need to ensure ethical diligence in child-centred technologies is more important than ever. From educational platforms like Duolingo ABC and Scratch to mental health companions like Woebot and Replika, AI is being utilized to support learning, creativity, and emotional well-being. However, as these tools expand in influence and reach, they raise critical ethical questions regarding the collection of children's data, the decision-making processes of algorithms, and the safeguarding of children's rights within digital environments.

As children engage with AI-powered technologies designed for their learning and development, the collection and use of their data raises critical ethical concerns. Most of the time, neither children nor their guardians are aware of how their data is collected, stored, or used, which poses a threat to transparency and informed consent. In the context of child-centred technologies, these issues are heightened by the need to protect vulnerable users' privacy and autonomy. In addition, algorithmic decisions can unintentionally perpetuate biases, resulting in unfair treatment or reinforcing harmful stereotypes, especially for children from diverse groups (Kannan, 2024; Gándara et al., 2024). To address these ethical challenges, careful monitoring, supervision, inclusive design practices, and policies that prioritise children's rights and well-being in the digital environment are required.

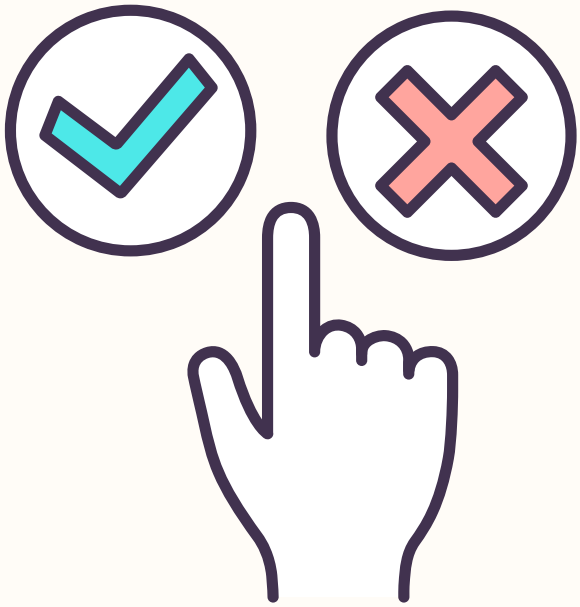
Technology keeps playing a significant role in children's everyday lives, but this increased exposure also makes them more vulnerable. Many children do not fully understand what personal information they are sharing or how it might be used. Without the right support or tools, they cannot always give informed consent. At the same time, many AI systems are not transparent, which can lead to privacy violations, unfair treatment, or design choices that take advantage of children's emotions and attention spans (Achieng, 2023; Lieber, 2018).

Accordingly, scholars and global organisations such as UNICEF (2021) and the Berkman Klein Centre at Harvard (Gasser et al., 2020) continue to emphasise the need for ethical frameworks specifically designed with children in mind. These frameworks should be grounded in core principles such as transparency, accountability, inclusivity, and a deep respect for children's autonomy. Ethical AI in child-centred technologies must prioritise privacy and minimise intrusive data collection, while ensuring that tools are age-appropriate, easy to understand, and inclusive by design. Furthermore, the development and governance of these technologies should actively involve educators, parents, policymakers, and developers, fostering shared responsibility for ethical oversight. Maintaining accountability through transparent governance frameworks and enforceable ethical standards is essential to safeguard children's rights in the evolving landscape of technological innovation.

There are successful models, such as Learning Passport and Canopy, that demonstrate how AI can be used ethically in child-centred technologies. This white paper draws on global research, case studies, and lessons from failures to propose actionable guidelines for child-centred AI, ensuring it empowers young users safely and equitably. It explores the ethical perspectives of AI and data use in child-centred technologies, focusing on how systems can be designed to protect children's rights, minimise risks, and promote their well-being to help them grow and thrive in the digital world.



KEY ETHICAL CONCERNS



According to a study by Berson et al. (2025), the rapid integration of artificial intelligence (AI) into early childhood education (ECE) presents transformative possibilities. However, it also raises urgent ethical challenges that demand immediate attention. Their study revealed significant gaps in safeguarding children's sensitive data, with inadequate protections against breaches, profiling, and misuse. Emotional AI tools, such as social robots and emotion-recognition technologies, offer novel learning opportunities but also risk undermining relational learning and fostering overreliance, manipulation, or loss of autonomy. The lack of developmentally appropriate design in AI systems further worsens these risks by failing to align technological solutions with the unique needs of young learners. Algorithmic bias, driven by non-representative datasets, perpetuates systemic inequities by disproportionately affecting marginalised communities and eroding fairness.

Regulatory frameworks appear fragmented and inconsistent, often lacking provisions tailored to the vulnerabilities of children or mechanisms for global enforcement (Berson et al., 2025). The integration of artificial intelligence (AI) into child-centred technologies presents issues related to data and privacy protection, algorithmic bias and its effects on children, and the potential for increased surveillance.

Data and Privacy Protection

Today, children's exposure to AI-driven technologies through educational tools, toys, apps, and smart devices is on the rise. These platforms often require access to sensitive personal data, including names, location, facial recognition, voice recordings, behaviour patterns, and emotions (Simas Velez, 2023; UNICEF, 2021). Children lack the cognitive maturity to understand data usage, storage, or sharing, and this raises ethical concerns.

Key Concerns

- **Informed Consent and Autonomy**: Children are not legally or developmentally capable of giving informed consent. Many decisions are made by parents or guardians, who may not fully comprehend the data collection mechanisms present in child-focused technologies (UNICEF, 2021; Achieng, 2023).

- **Data Exploitation and Commercial Use**: In some cases, children's data is utilised for advertising or commercial profiling purposes. This raises important questions regarding data usage, especially in situations where children are prompted to make in-app purchases or engage with content based on the collected information (Goodin, 2024).

- **Security and Data Breaches**: Neglecting data security can expose children's personal information to hacks, leaks, or unauthorised access. Given their long future ahead, children are particularly vulnerable to the long-term consequences of such breaches, such as identity theft and manipulation (Simas Velez, 2023; Achieng, 2023).

AI-powered devices and applications designed for children often collect extensive personal data, including names, locations, behavioural patterns, and even biometric information. This data collection poses significant privacy risks, especially given children's limited understanding of data usage and consent. Unauthorised access or data breaches can lead to misuse of sensitive information, potentially harming children's safety and well-being. Moreover, the opacity surrounding data collection, storage, and sharing exacerbates these issues, causing parents and guardians to question the security of their children's information (Simas Velez, 2023).

Algorithmic Bias and Its Impact on Children

Algorithmic bias occurs when AI systems reflect and perpetuate societal prejudices due to skewed or incomplete training data. These biases can have serious implications for children, especially in settings like education, healthcare, and content moderation (Berson et al., 2025; Goodin, 2024).

Key Concerns

- **Unfair Educational Outcomes:** Adaptive learning platforms or AI tutors that use biased data may misinterpret children's learning abilities or provide inadequate support based on gender, ethnicity, or socioeconomic background (Goodin, 2024; Berson et al., 2025).

- **Marginalisation and Reinforced Stereotypes:** AI systems that filter content or provide recommendations may reinforce harmful stereotypes, limiting exposure to diverse perspectives.

For example, children from different communities may have varied representations in AI-generated content (Achieng, 2023). What this means is that if child-centred AI tools are built using data that ignores one continent or region (e.g. Africa), the resulting content will not feature the cultures and realities of that continent and will reflect only a narrow view of childhood, leaving out children from remote villages, nomadic communities, informal settlements etc.

- **Digital Exclusion:** Non-inclusive AI technologies may exclude children with disabilities, neurodivergent traits, or those from underrepresented groups (UNICEF, 2021). AI systems are susceptible to biases present in their training data, which can result in unfair or discriminatory outcomes. In child-centred technologies, such biases may reinforce stereotypes or inadvertently marginalise certain groups of children based on race, gender, or socioeconomic status. For instance, biased educational AI tools might favour students from particular backgrounds, leading to unequal learning opportunities and outcomes. A study by researchers at Stanford University revealed that large language models (LLMs) used in educational platforms often reinforce harmful stereotypes and portray marginalised groups negatively in AI-generated educational content (Kannan, 2024).

In the same vein, research on AI models that predict student success found that there were racial disparities, as these AI tools were less accurate for racially minor students, which had the potential of leading to misinformed academic support decisions (Gandara et al., 2024). Additionally, tools that function as AI-driven plagiarism detection systems have been shown to tend to disproportionately flag black students due to misinterpretation of dialect and linguistic style, which further reinforces bias and mistrust in the system (Hirsch, 2024).

Beyond the bias in the algorithmic design of AI tools, inequitable access to power and educational tools also deepens inequality; students from underprivileged backgrounds often face barriers to accessing the digital infrastructure required to benefit from these innovations (Al-Zahrani, 2024). Addressing these biases is crucial to ensure that AI applications promote inclusivity and fairness among all children (Goodin, 2024; Berson et al., 2025).

The Risk of Over-Surveillance

AI technologies enable real-time surveillance of children's online and offline behaviour, locations, emotions, and activities. Though often justified by safety concerns like cyberbullying and school shootings, these capabilities raise significant ethical and psychological issues (Simas Velez, 2023; UNICEF, 2021).

Key Concerns

- **Invasion of Privacy and Trust Erosion:** Constant monitoring through AI-enabled cameras, facial recognition, and activity trackers may erode children's sense of privacy and personal space. When children feel constantly watched, it may alter their behaviour and inhibit their development of independence (Simas Velez, 2023).

- **Normalising Surveillance Culture**: Excessive surveillance in schools and homes may normalise constant monitoring. This undermines the value of privacy and autonomy as children grow up (UNICEF, 2021; Achieng, 2023).

- **Psychological Impact:** Studies suggest that being subject to surveillance from an early age can increase anxiety, reduce trust, and diminish creativity and spontaneity in children (Berson et al., 2025; Simas Velez, 2023).

While the intention of surveillance may be to ensure safety, excessive monitoring can infringe on children's privacy rights and hinder their autonomy. For instance, the implementation of AI-powered surveillance tools in educational institutions to monitor student behaviour has faced criticism. Concerns have been raised that such measures could foster a surveillance culture, wherein students may feel perpetually observed and assessed, and this kind of environment can negatively affect their development and well-being (Simas Velez, 2023).

Therefore, these ethical concerns need to be addressed. It requires a collaborative effort among policymakers, educators, technology developers, and parents to set up robust data protection laws, implement unbiased AI algorithms, and create guidelines that balance children's safety with their right to privacy. Ensuring transparency, accountability, and inclusivity in the design and deployment of AI technologies is essential to safeguard children's rights in the digital age.

This can include:

- Strong Regulatory Frameworks like GDPR (General Data Protection Regulation) and COPPA (Children's Online Privacy Protection Act) to enforce data protection and privacy rights for children (UNICEF, 2021).
- Leveraging ethical AI design principles, such as fairness, accountability, and transparency, when creating technologies aimed at children (Goodin, 2024; Berson et al., 2025).
- Education and awareness for parents, teachers, and children about how AI systems work and the risks they may pose (Achieng, 2023).
- Child participation in design, where children take part in shaping the technologies intended for their use, respecting their evolving abilities and rights (UNICEF, 2021).

GUIDELINES FOR ETHICAL AI DESIGN IN CHILD-CENTRED TECHNOLOGIES

Artificial intelligence is transforming how children learn, play, and interact in digital spaces. Unlike adults, children are not just smaller users; they are growing minds navigating a world they are learning to trust. While AI offers significant potential to enhance education and safety, it poses unique risks to children, whose developing minds are vulnerable to biased algorithms, privacy breaches, harmful emergent behaviours, or harmful content. For example, a tutor app or a learning app that starts nudging a child toward harmful content, or a chatbot that subtly reinforces harmful stereotypes. This is not science fiction; it is the reality of today's world. We have seen it in cases like the 2019 TikTok Children's Online Privacy Protection Act (COPPA) violation, where unauthorised data collection from children under 13 led to a $5.7 million fine (Yu et al., 2024).

How then do we shield children in a world where machines "think" in ways we cannot always predict? The answer to this is not to be afraid of progress but to shape it. To harness AI's benefits while protecting children, ethical designs must prioritise transparency, privacy, fairness, age-appropriate interaction, and inclusive and human-centred development. Jobin, Lenca, and Vayena (2019) noted that in the past five years, private companies, research institutions, and public sector organisations have issued principles and guidelines for ethical AI. To reduce harm and enhance benefits, AI systems must be in line with these basic moral principles, social standards, and legal frameworks. The fundamental rules of ethical AI design are examined in this section, and it covers concepts like accountability, openness, fairness, and privacy.

Guidelines for Ethical AI Design

To create AI systems that prioritise child safety and well-being, developers must adhere to four key principles: transparency and accountability, privacy and autonomy, age-appropriate design, and collaborative policy development. These guidelines, grounded in research and real-world application, provide a framework for ethical AI deployment in child-centric environments.

1. Ensuring Transparency and Accountability in AI Systems for Children

Transparency and accountability are essential components of AI and Data ethics in child-centred technologies. AI systems must be transparent about their decision-making processes and accountable for their outcomes, especially when used by children.

Transparency means clearly explaining how AI works, what data it uses, and how decisions are made, in ways accessible to parents, educators, and regulators. Floridi and Cowls (2019) propose explicability as a core ethical principle, combining intelligibility (how does it work) and accountability (who is responsible). For example, an AI recommending educational content should disclose its logic, such as basing suggestions on a child's learning history (Jobin et al., 2019). The European Commission's Artificial Intelligence Act (2021) mandates such transparency for high-risk systems in education, requiring clear disclosures about data usage and decision rationales.

To enhance transparency, developers can implement "glass box" models, logging data interactions for real-time auditing, as proposed by Jobin et al. (2019). Miller (2019) suggests using age-appropriate language and visuals to explain AI actions to children, such as why a game recommends specific activities. Accountability requires developers to take responsibility for negative impacts, such as biased outcomes, through regular audits and public reports (Brundage et al., 2020). For instance, audit trails and third-party certifications can verify compliance with privacy laws like COPPA and GDPR-K. A practical example is an AI-driven educational tool providing parents with a dashboard showing data usage and decision rationales, fostering trust and accountability.

Relating this to developing child-centred technologies, products developed for children leveraging AI must ensure transparency and a system of accountability. Children should be notified in a forthright manner when they interact directly with an AI system, to avoid a situation where they believe they are interacting with a human (UNICEF, 2021). This is critical to ensure child safety and reduce unintentional harm to children.

Strategies for Ensuring Transparency

- **Explainability and Interpretability:** AI models, particularly complex ones like deep learning algorithms, must be designed to provide explanations for their decisions. Explainable AI (XAI) techniques help make AI decision-making processes more interpretable to users and regulators (Miller, 2019). A good example is Carnegie Learning's MATHia platform. "MATHia" is an adaptive learning platform that helps children learn mathematics by personalising instruction. It is used widely in schools in the United States and is designed specifically with students in mind (Carnegie Learning, 2024).

- **Open Documentation and Reporting:** Developers should maintain detailed documentation on AI training data, algorithms, and decision-making processes. Reports on AI operations that are made publicly available foster trust and permit external audits (Brundage et al., 2020). An example of open documentation and reporting is the Model Card introduced by Google. Model Cards are a documentation framework introduced by Google AI to promote transparency, accountability, and ethical deployment of machine learning models. They serve as structured summaries of key facts about a trained model, enabling users, whether developers, policymakers, or end users, to understand what the model does, how it was trained and evaluated, and where and how it should or should not be used (Google AI, 2025).

- **Stakeholder Engagement:** Rahwan et al. (2019) suggested that AI designers should involve various stakeholders, including ethicists, policymakers, and end-users, in the development process to ensure that transparency and accountability concerns are addressed from multiple perspectives.

Strategies for Ensuring Accountability

- **Clear Legal and Ethical Guidelines**: Governments and organisations should establish clear policies and regulations that define responsibilities in AI development and deployment. Ethical guidelines such as the EU's AI Act emphasise accountability measures for AI systems (European Commission, 2021). The EU emphasises that when

developing child-centred technologies, particularly those that leverage artificial intelligence (AI), developers must follow ethical, legal, and safety guidelines to ensure children's rights, development, and well-being are protected. According to UNICEF (2021), these guidelines address areas of concern such as the child's interests, age-appropriate design, confidentiality and explainability, privacy and data protection, safety and security, fairness and non-discrimination, informed consent, agency, empowerment, and inclusion.

- **Human Oversight and Auditing:** AI systems should include mechanisms for human oversight to prevent unethical or harmful decisions. Regular audits can help detect biases, errors, and unintended consequences (Jobin et al., 2019).

- **Liability Frameworks:** Organisations deploying AI should have clear liability frameworks outlining who is responsible for AI-related harm. These can include developers, users, or companies that implement AI technologies (Russell & Norvig, 2020).

> AI systems must be designed to respect children's rights by ensuring secure data handling, transparency, and mechanisms that empower them with greater control over their personal information.

2. Prioritising Children's Privacy and Autonomy

Children are particularly vulnerable to data exploitation due to their limited understanding of digital privacy risks and their reliance on online services (Livingstone & Stoilova, 2021). Children's vulnerability demands robust privacy protections and designs that promote autonomy within safe boundaries. AI systems must be designed to respect children's rights by ensuring secure data handling, transparency, and mechanisms that empower them with greater control over their personal information. Therefore, child rights should be integrated into data science, emphasising minimal data collection and strong encryption to prevent breaches (Berman and Albright, 2017). The COPPA and GDPR-K require anonymisation and robust encryption to protect children's data.

To ensure children's privacy, we must consider the following:

- **Minimising Data Collection:** AI systems should only collect the minimum amount of data necessary for functionality. Excessive data collection increases the risk of privacy violations and potential misuse. Age-verification systems and parental consent protocols that align with COPPA are recommended to restrict data collection for users under 13 (UNICEF, 2021). An example is the UNICEF project on good governance of children's data, which maintains that every child is different, with unique identities, and their capacities and circumstances evolve over their life cycle. Children are more vulnerable than adults and are less able to understand the long-term implications of consenting to their data collection. For these reasons, children's data deserves to be treated differently and securely.

- **Secure Data Storage and Processing**: Organisations must implement robust encryption and anonymisation techniques to protect children's data from unauthorised access, cyber threats, and breaches.

- **Parental and Guardian Oversight:** There is a gap in parental awareness of children's generative AI (GAI) use, such as emotional interactions with chatbots. These inadequate parental controls force reliance on manual checks (Yu et al., 2024). While children should have some autonomy over their data, appropriate parental oversight is essential. AI platforms should incorporate parental control features while balancing children's right to digital privacy. An example is Canopy, an AI-driven application that provides real-time protection for children navigating the internet. Unlike traditional parental control apps that rely solely on predefined lists of blocked websites, Canopy utilises advanced AI to analyse and filter content dynamically, ensuring that inappropriate material is identified and blocked as it appears (Canopy, 2025).

Consequently, autonomy-safe designs that allow children to make choices within predefined boundaries, such as selecting a game level without manipulative nudging, are recommended (Gasser et al., 2020).

For example, a storytelling AI should enable parents to audit and limit data shared during interactions, using end-to-end encryption and granular privacy controls to ensure safety.

- **Clear and Child-Friendly Privacy Policies:** Privacy policies must be written in a language that children can understand, allowing them to make informed choices about their data (Livingstone & Stoilova, 2021). In addition, these policies must also provide a framework that can protect the children.

To promote children's autonomy, we must also consider the following:

- **Informed Consent Mechanisms:** AI platforms should provide children with age-appropriate explanations of how their data is used and obtain meaningful consent before data collection (UNICEF, 2021). For example, in this report, UNICEF maintained that a child who interacts directly with an AI system (e.g. a toy, chatbot, or online system) has the right to receive an explanation at an age-appropriate level and inclusive manner, including through the use of animations, to understand how the system works and how it uses and maintains data about them. Requirements of explanation, transparency, and redress also apply to AI systems that impact children indirectly.

- **Right to Data Deletion**: According to the UK Information Commissioner's Office (ICO) (2025), guidance on children's data rights, leveraging AI systems designed for children must allow children and their guardians to review and delete personal data upon request especially if it seems likely that they gave their data without fully understanding the implications of doing so. This right is crucial in maintaining children's autonomy in digital spaces.

- **Ethical AI Design for Child-Centred Decision-Making:** Lieber (2018), a professional psychologist, lamented about how tech companies manipulate human behaviour. They believe that psychology is being used as "a weapon against children". This is the reason child-centred tools leveraging AI should empower children by giving them control over their digital interactions rather than manipulating their behaviour through persuasive design techniques.

3. Developing Age-Appropriate Tools

AI systems must align with children's cognitive and emotional development, using age-appropriate language, visuals, and interactions. Therefore, it is important to leverage UI/UX best practices such as intuitive interfaces, content filters to block harmful material, and feedback mechanisms for children to report issues (Ragone et al., 2024). However, we must exercise caution against addictive features like excessive gamification in conversational AI (CAI), which can exploit children's attention (Chubba et al., 2021).

UNICEF (2021) emphasised content moderation using AI trained on child-safe datasets to filter inappropriate content. Developers need to adapt tools to a child's learning pace and emotional state, thereby providing supportive feedback. User testing with diverse age groups ensures interfaces are intuitive and safe. For example, an AI educational game for young children could use bright, simple visuals and robust content filters, while older children might engage with interactive quizzes tailored to their skill level.

Key Considerations in Developing Age-Appropriate AI Tools

- **Cognitive and Emotional Development:** AI systems should be designed with an understanding of children's cognitive and emotional growth stages. For younger children, AI should use simple, interactive interfaces with minimal complexity, while for older children, more advanced AI tools can offer personalised learning and problem-solving features (Holloway, Green, & Livingstone, 2021).

- **Adaptive Learning and Personalisation:** AI-driven educational platforms should offer personalised learning experiences based on a child's progress, interests, and learning style. However, personalisation should not compromise children's autonomy or manipulate their behaviour through algorithmic nudging (Blanchard et al., 2021). The impact of this manipulative behaviour through algorithmic nudging could be seen in an app which was designed to show age-appropriate content to children. The platform used AI algorithms to recommend videos based on what the child had previously watched, such as educational cartoon videos. After watching, the AI recommended another similar video, not based on educational value or the child's actual intention but based on what kept the child watching. This scenario loops a child into watching low-quality or sensational content (e.g., unboxing videos, repetitive cartoons, etc). The child may end up spending hours watching videos, not because he or she chose to, but because the system continuously nudges them with content tailored to exploit their preferences.

- **Child-Friendly User Interfaces:** AI tools should be designed with age-appropriate language, visual elements, and voice-based interactions that align with children's developmental stages. Ensuring that interfaces are engaging yet simple helps prevent frustration and enhances usability.

- **Ethical Data Collection and Privacy Protections**: AI tools must comply with strict privacy standards by minimising data collection, anonymising user information, and incorporating parental consent mechanisms. Age-appropriate AI should also educate children about digital privacy in ways they can understand (UNICEF, 2021).

- **Protection from Harmful Content:** AI-powered platforms, particularly in social media and gaming, must have safeguards to protect children from harmful content, cyberbullying, and inappropriate interactions. AI moderation should align with child protection policies and be transparent in its decision-making processes (Livingstone & Stoilova, 2021). During an academic roundtable for the Global Online Safety Regulators Network hosted by the Digital Future for Children (DFC), eSafety Commissioner Julie Inman Grant stressed that children are at risk of being exposed to various harms in most digital environments. Even if 'algorithmic perfection' were possible, a risk of harm would still exist in social media environments as platforms reflect their users' behaviours and their underlying business models. At this roundtable, Snapchat was highlighted as problematic because it had little effective age verification and 'mass adding' functions. Therefore, it posed further monitoring issues too, as it operated in the balance between a social media and messaging platform. (Grant, 2024).

- **Encouraging Creativity and Critical Thinking:** Child-centred AI tools should encourage children's creativity and critical thinking instead of simply automating tasks. Tools such as StoryWizard.ai, an AI tool that helps children create stories by suggesting plot ideas, dialogue, and illustrations, and Scratch (by MIT Media Lab), a visual programming language for children to create games, animations, and simulations, can foster innovation and independent problem-solving skills (Blanchard et al., 2021).

4. Collaborative Policy Development

Engaging parents, educators, children, and child-focused organisations in AI design and testing ensures diverse perspectives shape ethical solutions. Gasser et al. (2020) highlight the value of participatory design, citing UNICEF's collaboration with children to develop its AI policy as a model. This approach fosters trust and aligns systems with real-world needs. For instance, involving children in testing a storytelling AI can ensure its language and content resonate with their developmental stages. Collaborative frameworks also help establish child-friendly consent mechanisms and privacy settings, as recommended by Berman and Albright (2017). Developers should conduct workshops with stakeholders to co-create guidelines, ensuring AI reflects the needs of its young users and their advocates.

Case Studies of Ethical AI Design

The following examples illustrate how the proposed guidelines are applied in practice, showcasing ethical AI systems that prioritise child safety and well-being.

Tool	Transparency	Privacy	Age Appropriate Design	Collaboration
Ask Save the Children (2024)	Provides educators and parents with clear explanations of AI-driven advice, supported by audit logs.	Uses anonymised data and minimal collection, adhering to GDPR.	Tailors advice to age-specific needs, ensuring accessibility	Engages child protection experts in design
UNICEF AI for Child Safety (2021)	Publishes transparent reports on AI performance, meeting EU AI Act standards	Employs anonymised data and encryption, complying with COPPA and GDPR-K	Filters content to ensure age-appropriate interactions, using child-safe datasets	Collaborates with global child rights advocates to refine systems
Collaborative Game Design Project (2021)	Shares design rationales with students and educators, ensuring clarity.	Collects minimal data, prioritising student privacy	Uses age-appropriate interfaces to foster engagement	Involves students in co-design, promoting autonomy
Edves (Nigeria)	Offers transparent tracking of student progress and AI interventions	Complies with the Nigerian Data Protection Act through secure data practices	Delivers self-paced, age-appropriate modules	Partners with schools and governments for localised solutions

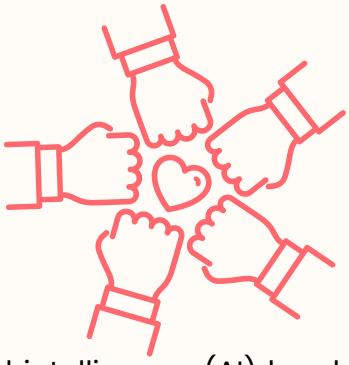
Table 1: Case Studies of Ethical AI Design

> ...developing age-appropriate AI tools is vital in promoting children's cognitive, emotional, and social development.

In conclusion, it is important to consider transparency, accountability, children's privacy protection, autonomy, and age-appropriateness when developing child-centred technologies. Ethical design of AI systems, particularly those interacting with children, requires a comprehensive approach that ensures transparency, accountability, privacy, and age-appropriate development. Transparency and accountability are crucial for fostering trust in AI technologies by making decision-making processes understandable and holding responsible entities liable for their actions. Strategies such as explainability, stakeholder engagement, and regulatory frameworks help ensure that AI systems operate in a fair and just manner. Prioritising children's privacy and autonomy is essential in protecting them from data exploitation and ensuring their rights are respected. AI systems should implement strong data protection measures, provide child-friendly privacy policies, and offer mechanisms for informed consent and data control. This ensures that children can engage with AI technologies safely without compromising their personal information.

More importantly, developing age-appropriate AI tools is vital in promoting children's cognitive, emotional, and social development. AI-driven systems should be designed to align with children's evolving abilities, provide adaptive learning experiences, and encourage creativity while safeguarding them from harmful content. Ensuring ethical data collection, privacy protections, and interactive interfaces enhances the usability and effectiveness of AI tools for young users.

THE ROLE OF STAKEHOLDERS



Artificial intelligence (AI) has become a bigger part of our daily lives, and it is shaping how we learn, work, communicate, and even care for our health. It is vital to ensure these systems are designed ethically. But building responsible AI isn't just a job for tech developers and engineers. It is a shared responsibility that involves many different people, such as government policymakers, educators, researchers, industry leaders, parents, and even the everyday users of these technologies. Every stakeholder has a part to play in making sure AI is fair and respects human rights values, most especially the rights of children.

According to Kirkland and Tsinaraki (2025), ethical concerns surrounding the integration of AI in child-centred technology extend beyond developmental considerations to include issues such as data privacy, algorithmic bias, and accountability that cannot be solved in isolation; they require collaboration. When different voices and perspectives come together throughout the design and development process, the result is AI that better serves society and protects those who are most vulnerable. Understanding the role each stakeholder plays is the first step toward creating AI that is not only smart but also safe and trustworthy.

The Role of Tech Companies

AI advancements keep transforming modern life, with tech companies playing a pivotal role in their development and implementation. These firms are not just the architects of AI innovations but also the key decision-makers in how these technologies are applied across society. Given their influential position, tech companies such as Microsoft, in collaboration with UNICEF, have worked together to create principles for child-centred AI, bearing a major responsibility in establishing and maintaining ethical guidelines. (UNICEF, 2021). Adhering to best practices, Microsoft, Google (YouTube for Kids), and Apple have been identified as tech companies with best practices on AI design for children. Their principles emphasise privacy, transparency, and avoiding unfair bias for children (Google, 2018).

To further enhance data security, Google has updated YouTube Kids app and included enhanced data safety disclosures, noting that no data is shared with third parties, browsing history is encrypted in transit and children's data can be deleted upon request. In addition, the newly launched Google Gemini AI, can be accessed by children under 13 with strict parental supervision. No child data was used in model training and full parental oversight was enabled (Google, 2025). Although government oversight is crucial, the rapid evolution of AI demands that these companies also take proactive steps in self-regulation and internal governance to promote responsible development and deployment. They must ensure the following measures are put in place:

a. Ethical Design and Development for Child-Centred Technologies:
One of the most critical responsibilities of tech companies is to embed ethical principles directly into the design and development of child-centred technologies. This includes ensuring that systems are transparent, fair, inclusive, and respectful of user privacy. Companies such as Google, Microsoft, and IBM introduced internal AI ethics boards and published guiding principles for responsible AI use. These frameworks often emphasise core values such as fairness, transparency, safety, and accountability (Floridi & Cowls, 2019).

To act responsibly, companies need to assess the potential impact of their AI technologies before deployment. This means conducting ethical risk assessments, bias audits, and stakeholder consultations to identify and mitigate possible harms. According to Jobin, Lenca, and Vayena (2019), this proactive approach not only helps prevent misuse but also builds public trust in emerging technologies.

b. Transparency and Communication
Tech companies must also lead by example in making their AI systems more understandable to both regulators and the public, especially children. Open communication about how algorithms work, what data they use, and what outcomes they produce is vital. Companies can publish documentation, open-source parts of their models, or offer explainable AI features that help users understand why a system made a particular decision (Miller, 2019).

c. Collaboration with Regulators and Civil Society
While self-regulation is important, tech companies should not operate in a vacuum. Instead, they should actively collaborate with governments, academia, and civil society groups to shape regulations that are both practical and protective. Initiatives such as the Partnership on AI and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems demonstrate how industry-wide cooperation can lead to meaningful guidelines and best practices. Furthermore, companies have the resources and expertise to support the development of global AI standards, particularly in areas like algorithmic fairness, safety protocols, and children's digital rights. By partnering with independent experts and advocacy organisations, they can ensure their technologies are not just profitable but also aligned with societal values.

d. Internal Governance and Culture
Tech companies should create a culture of ethical responsibility within tech companies. This includes training staff in AI ethics, encouraging whistleblowing, and establishing clear internal policies that promote responsible innovation. Having dedicated teams for AI ethics and ensuring those teams have real influence in product development decisions is key to embedding ethics into the company's culture.

The Role of Governments

The rapid integration of AI into healthcare, education, finance, security, transportation, etc., demands strong governmental oversight. While tech companies innovate, governments must regulate, standardise, and protect societal interests. Without proper governance, AI risks undermining fairness, accountability, privacy, and safety, making public-sector leadership indispensable in upholding ethical standards and human rights.

a. Establishing Legal and Regulatory Frameworks for Child-Centred Technology

One of the most direct ways governments influence AI is by creating laws and regulations that govern its use. These include general data protection regulations, algorithmic transparency requirements, and frameworks for AI safety. For instance, the European Commission's proposed Artificial Intelligence Act is one of the most comprehensive legal attempts to regulate AI, classifying AI systems based on risk and setting obligations for each category (European Commission, 2021). Such legal frameworks serve to prevent misuse, ensure user rights, and establish consequences for harm caused by AI systems. Governments also define ethical and legal boundaries for specific AI applications such as facial recognition, predictive policing, or autonomous weapons. By doing so, they provide clear signals to industry on acceptable and unacceptable practices.

b. Promoting Transparency and Accountability

Governments have the authority to require AI systems to be explainable and auditable, particularly in critical areas like criminal justice, healthcare, and education. For example, public agencies may mandate algorithmic impact assessments, much like environmental assessments, to evaluate potential harm or bias in automated systems (Doshi-Velez et al., 2017). Additionally, governments can establish independent oversight bodies to monitor AI deployment and investigate complaints relating to child-centred technologies leveraging AI. These institutions can help maintain public trust and ensure accountability when AI technologies affect children's lives.

c. Protecting Data Privacy and Civil Liberties

AI systems rely heavily on data, including personal and sometimes sensitive information. It is the role of governments to ensure that data is collected, stored, and used responsibly. Legal instruments such as the General Data Protection Regulation (GDPR) in Europe are examples of how governments can protect citizens' rights by giving them control over their data, requiring consent, and enforcing penalties for violations (Floridi & Cowls, 2019). In democratic societies, governments are expected to safeguard civil liberties by preventing intrusive surveillance and discriminatory outcomes from AI systems.

Regulations that limit biometric surveillance or require public approval before deploying certain AI tools exemplify this responsibility. Governments should ensure the privacy and security of data in children-centred technologies to boost users' confidence in using such technologies.

d. Ensuring Inclusivity and Equity

Governments must ensure that AI benefits all members of society, including marginalised and vulnerable populations. This includes addressing digital divides, promoting diverse participation in AI design, and implementing inclusive procurement policies. Public policy can incentivise ethical business practices and discourage companies from deploying biased or exploitative AI systems. Governments also have a role in public education and awareness. By informing citizens about their rights and the impact of AI, governments empower people to make informed choices and hold institutions accountable.

Children who are already marginalised due to poverty, geographic location, disabilities, or systemic inequality face a double burden when it comes to accessing AI-leveraged technology. First, they are excluded from the potential benefits AI can bring in education, healthcare, and social services. Second, they are more likely to be overlooked in the design and policymaking processes that shape these technologies. Governments must proactively address the digital divide by investing in infrastructure (e.g., affordable internet, devices, and electricity) in underserved areas. Ensuring inclusive design where AI educational tools, health apps, and learning platforms are not just built for wealthy or urban populations but also adapted for rural, low-income, and differently abled children. Implementing policies that prioritise funding for schools and community centres that serve marginalised children to access and use AI technologies meaningfully.

A good example of such an initiative is Giga, an initiative that aims at connecting every school in the world to the Internet and, by extension, every young person to information, opportunity, and choice. This project focuses on closing the digital divide, especially for children in low-income and rural communities who otherwise would have no access to AI-powered educational resources. (UNICEF, 2023).

The Role of Educators and Parents

As AI increasingly becomes part of children's daily experiences, from virtual assistants and personalised learning tools to entertainment apps, parents and educators hold a key responsibility in guiding young users' engagement with these technologies. Their influence is essential in helping children build safe, ethical, and purposeful interactions with AI. By promoting digital literacy, safeguarding emotional health, and advocating for responsible AI use, these stakeholders can shape a future where technology supports children's growth and well-being.

a. Educators as Facilitators of AI Literacy

Teachers and school administrators are on the front line of introducing AI technologies into the classroom. Their role is not just to use these tools effectively but also to educate students about how AI works, its limitations, and its ethical implications. Introducing children to basic AI concepts, such as how machines learn, what data they use, and how biases can form, can help demystify the technology and build critical thinking skills. For instance, educators can teach students to question algorithmic decisions or recognise when a recommendation system may not be fair. This empowers children to be more thoughtful consumers and creators in a digital world.

Furthermore, teachers must stay informed about the types of AI-enabled tools their schools use, ensuring they are age-appropriate and aligned with educational goals. Collaborating with policymakers, IT specialists, and child psychologists can also help in selecting or designing AI systems that support learning without undermining privacy or autonomy.

b. Parents as Digital Mentors and Protectors

At home, parents play a crucial role in setting boundaries and modelling healthy digital behaviour. Many children encounter AI through voice-activated toys, recommendation systems on streaming services, or chatbots embedded in apps. Without guidance, these interactions can lead to overdependence, privacy breaches, or exposure to inappropriate content. Parents should engage in open conversations with their children about how AI systems work and what information they collect. Encouraging questions like "Why is this app suggesting this video?" or "What do you think this robot knows about you?" help children become more reflective users. Also, parents should monitor the types of AI applications their children use, review privacy settings, and ensure compliance with regulations like COPPA (Children's Online Privacy Protection Act) or GDPR-K (General Data Protection Regulation for Kids)

c. Promoting Emotional and Social Development

AI technologies can sometimes create the illusion of companionship or authority, especially with conversational agents and learning assistants. Educators and parents must help children distinguish between human relationships and machine interactions. This includes explaining that AI systems do not have emotions or intentions, even if they appear friendly or understanding. Adults should also be mindful of how AI may affect children's self-esteem, attention spans, and interpersonal skills. For example, overreliance on AI tutors could reduce opportunities for collaborative learning and emotional support from real teachers or peers (Holloway et al., 2021).

d. Advocacy and Participation in Design

Educators and parents must also advocate for ethical AI design by voicing their concerns to developers, school boards, and policymakers. Their insights into children's needs, behaviours, and developmental stages are invaluable during the design and testing phases of AI tools. Some initiatives involve "participatory design," where parents, teachers, and even children contribute to the creation of AI systems. A research

project funded by UK Research and Innovation named "Fair-AIEd Participatory Design" is a great example of this practice. It was an initiative that involved creating ethical AI systems for educational contexts through participatory research. This initiative sought to involve various stakeholders, including educators and potentially students, to ensure that AI applications in education are fair and inclusive. According to UK Research and Innovation (2021), the impacts of AI technologies will be listed, mapped, and analysed to illustrate key issues and concerns in this emerging landscape and to identify potential for positive educational change. Upon identification of benefits, harms, and risks as they apply to AIEd, stakeholders will then co-design and develop a Fair-AIEd Trust Mark.

Key Stakeholders

Tech Companies

They are not just the architects of AI innovations but also the key decision-makers in how these technologies are applied across society.

Government

Governments must regulate, standardise, and protect societal interests. Without proper governance, AI risks undermining fairness, accountability, privacy, and safety.

Parents and Educators

They hold a key responsibility in guiding young users' engagement with these technologies. Their influence is essential in helping children build safe, ethical, and purposeful interactions with AI.

GOOD PRACTICE: SUCCESSFUL IMPLEMENTATION IN CHILD EDUCATION OR SAFETY

As artificial intelligence (AI) becomes more deeply embedded in digital environments, some developers and researchers have created ethical AI tools specifically tailored to enhance children's learning and protection. These solutions are grounded in key ethical principles, including transparency, developmental appropriateness, inclusivity, and robust privacy safeguards, making them particularly well-suited for young users.

The following examples illustrate some notable implementations of these child-centred AI technologies:

1. Duolingo Kids
Duolingo Kids is a language-learning app that uses AI to personalise content for young learners. The platform adapts to each child's learning pace, identifies areas of struggle, and provides encouraging feedback. Importantly, Duolingo ensures data minimisation and adheres to child privacy standards, making it one of the more ethically designed AI-powered educational apps (Duolingo, 2025).

Ethical Features: Personalised learning, minimal data collection, child-friendly interface, GDPR-compliant.

2. Woebot for Teens
Woebot is a mental health chatbot that uses natural language processing (NLP) to provide emotional support. A specialised version has been designed for teens, helping them navigate mental health challenges through guided conversations grounded in cognitive-behavioural therapy (CBT). While still experimental, Woebot emphasises privacy and emotional safety, offering support without collecting unnecessary personal data. (Woebot Health, 2024).

Ethical Features: Transparent AI interactions, minimal data retention, CBT-based, non-judgmental tone.

3. Scratch by "MIT Media Lab"
Scratch is a visual programming platform designed to introduce children to coding. Though not AI itself, it serves as a foundational tool for understanding how machines "think" and make decisions. It has been integrated with AI extensions that allow kids to experiment with basic machine learning concepts in an age-appropriate, creative way. (Scratch, 2025). MIT RAISE acknowledges Scratch's role in helping children learn to think creatively, reason systematically, and work collaboratively. The team noted that Scratch helps kids learn to think creatively, reason systematically, and work collaboratively. The organisation highlights Scratch's research into incorporating AI-based tools that

empower kids to create using speech recognition, body tracking, and object recognition, emphasising the platform's commitment to ethical AI integration (MIT Raise, 2025).

Ethical Features: Emphasis on education over surveillance, open access, community moderation, and no advertising.

4. Replika (Youth Mode)
Replika, originally developed for adults, has implemented a "Youth Mode" that restricts certain conversations and adds safety filters for younger users. It is a conversational AI that allows young users to explore their emotions and develop social skills in a controlled and privacy-conscious environment (Safer Schools, 2022)

Ethical Features: Adjustable settings for age-appropriate content, real-time moderation, and user data control.

5. UNICEF's Learning Passport
UNICEF developed a Learning Passport in partnership with Microsoft, an AI-driven platform designed to deliver quality education to children in crisis-affected regions. It uses AI to recommend learning paths and track progress, all while ensuring data privacy and accessibility in low-resource settings. (UNICEF 2024)

Ethical Features: Child protection focus, secure data architecture, and offline functionality.

6. The Adventures of Muna
The Adventures of Muna is a child safety education game app by The Safety Chic. It helps children learn about complex safety topics in a simple and memorable way. It ensures data minimisation and adheres to child privacy standards.

Ethical Features: Minimal data collection, child-friendly interface, GDPR-compliant, Firebase encryption

Lessons Learned from Breaches or Failures in Ethical Design

Despite several great examples of AI usage, the ethical challenges associated with its design and deployment have drawn significant attention. Several breaches in ethical AI design, especially those involving children, have demonstrated the high stakes of ignoring ethical principles such as fairness, transparency, accountability, and privacy. These failures serve as valuable case studies from which scholars, developers, policymakers, and educators can derive critical lessons for the future of responsible AI development.

1. Addressing Algorithmic Bias and Discrimination in Child-centred Technologies
One of the most recurring failures in ethical AI design relates to algorithmic bias, where AI systems reflect and reproduce societal inequalities embedded in training data. For instance, facial recognition software has been shown to perform disproportionately poorly on women

and individuals with darker skin tones due to underrepresentation in datasets (Buolamwini & Gebru, 2018). When such systems are used in child-centred technologies such as education, they risk perpetuating systemic inequalities in the educational system. Baker and Hawn (2022) highlight biases in educational AI that favour certain racial or socioeconomic groups due to flawed training data. For example, an AI recommending resources might exclude underserved students with disabilities if not inclusively trained and this could affect the mental health of affected populations.

Lesson learned: Regular audits of datasets and processes to ensure they reflect race, gender, disability, and socioeconomic status are essential to ensure equitable outcomes. It is crucial to incorporate diverse and representative datasets in the training process. Regular audits for fairness and bias mitigation techniques must be embedded in the AI development lifecycle (Mehrabi et al., 2021).

2. The Importance of Transparency and Explainability in Child-centred Technologies

A key issue in many ethical lapses is the opacity of AI systems, often described as "black boxes." A high-profile example is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). It is a risk assessment tool used in the U.S. criminal justice system to predict the likelihood that a defendant will "re-offend". It has been used in sentencing decisions, parole, and bail hearings. COMPAS's algorithm used in criminal sentencing has been found to produce racially biased outcomes without clear justifications for its decisions (Angwin et al., 2016). This example though not related to child-centred technologies provides a practical example of the importance of transparency and explainability.

Lesson learned: AI systems, especially those affecting vulnerable populations, must be interpretable. Developers should prioritise explainable AI (XAI) models and provide clear, user-friendly explanations of how decisions are made.

3. Ensuring Data Privacy and Security

AI systems that collect and process sensitive data, especially from children, pose significant privacy risks if not ethically designed. The 2015 VTech breach exposed over 6 million children's data due to poor encryption and excessive collection (Berman & Albright, 2017). Minimising data collection and implementing strong encryption, as mandated by GDPR-K, are vital to prevent breaches. Similarly, poorly secured educational apps have resulted in data leaks and unauthorised data exploitation. In addition, Yu et al. (2024) highlight parents' struggles to monitor children's GAI use, such as emotional interaction with chatbots, due to inadequate controls, as seen in real-world failures like Tik-Tok's 2019 COPPA violation, where the platform collected data from users under 13 without consent, leading to a $5.7 million fine.

Lesson learned: Designers must comply with data protection regulations and integrate privacy-by-design principles. For children, special care must be taken to gain verifiable parental consent, minimise data collection, and secure storage and transmission (UNICEF, 2021). Parental controls and age-verification systems are important. Therefore, AI systems for children must integrate robust parental controls and age verification systems to prevent unauthorised data use.

4. Avoiding Manipulative and Addictive Design

AI systems designed to maximise engagement can inadvertently or deliberately exploit psychological vulnerabilities. YouTube Kids and similar platforms have been criticised by Radesky et al. (2024) for using recommendation algorithms that expose children to excessive or inappropriate content in pursuit of watch time. Such designs can lead to digital addiction and cognitive overstimulation. However, according to Kidslox (2024) Youtube Kids has introduced stricter content filters, a more restrictive algorithm, and authoritative content prioritization. The system has implemented stronger filters to limit exposure to potentially harmful content; YouTube now prioritizes authoritative sources for children's educational content.

Lesson learned: Ethical design should emphasise well-being over profit. Developers should avoid addictive design patterns when building systems for children and instead focus on age-appropriate, time-bound, and educationally aligned content.

5. Establishing Clear Accountability Mechanisms and Human Oversight

Failures in ethical AI often reveal ambiguities in responsibility, whether among developers, deploying organisations, or regulatory bodies. When ethical breaches occur, victims are often left without recourse due to a lack of clear governance structures or enforcement mechanisms. Furthermore, Holloway et al (2021) observe that AI-powered platforms can amplify cyber-hate, with 20% of U.S. college students experiencing cyberbullying, as seen in YouTube Kids' 2017-2019 issues with inappropriate content recommendations.

Despite the study not being child-specific, it indicates the risk of AI content moderation failures. Therefore, robust content filters and human oversight are necessary to ensure safe, age-appropriate interactions.

Lesson learned: Accountability must be integral to AI governance. This includes designating responsible actors for decisions made by AI systems and creating transparent pathways for complaints, redress, and ethical oversight (Brundage et al., 2020).

6. The Necessity of Inclusive and Participatory Design

Finally, many ethical failures result from the exclusion of end users, particularly children, parents, and educators, from the design process. This top-down approach leads to tools that fail to align with user needs or cultural contexts. Ethically designed AI must include input from diverse stakeholders, especially those directly affected by the technology.

Lesson learned: Participatory design approaches foster trust and relevance. Engaging users in the early stages of development helps to identify potential risks, address usability concerns, and ensure cultural and contextual appropriateness (Gasser et al., 2020).

RECOMMENDATIONS: PATH TO CHILD SAFETY

In a world where most children are digital natives and comfortable with technology, it is imperative to ensure that technology they are exposed to complies with ethical guidelines that prioritise privacy, fairness, transparency, and accountability throughout their development and use.

The following are recommendations to help ensure safe AI usage and data ethics in child-centred technologies:

1. Incorporate Ethical Foundations in the Development of Child-Centred Technologies

To ensure that AI and data-driven technologies benefit children ethically and responsibly, it is important to establish strong ethical foundations from the very beginning. This involves integrating child-specific principles at every stage of development from initial design to final implementation. Frameworks such as UNICEF's Policy Guidance on AI for Children (2021) provide clear and consistent guidelines that should be applied throughout the development process.

2. Strengthen Data Protection and Privacy

Safeguarding children's data must be a top priority in the design and use of AI and digital technologies. This involves adhering to the principle of data minimisation by collecting only the strictly necessary information. Consent processes should be clear, age-appropriate, and easily understood by both children and their guardians. Crucially, children's rights must be protected, including the right to have their data erased and to withdraw consent at any time (ICO, 2025).

3. Proactively Identify and Mitigate Algorithmic Bias

To prevent unfair outcomes based on race, ability, gender, or other factors, AI systems designed for children must be carefully built and regularly tested to ensure they work fairly. When algorithms are biased, they can reinforce harmful stereotypes or limit opportunities for certain groups of children. Developers should use diverse and inclusive datasets, routinely check for bias, and involve independent reviewers to help assess fairness (Buolamwini & Gebru, 2018; Gándara et al., 2024). Promoting fairness in AI is essential to creating child-centred technologies that serve all children equally.

4. Promote Transparency and Explainability

Transparency and explainability are essential and should be prioritised by offering clear, simple explanations of how AI tools function and how decisions are made. These explanations should be accessible to parents, educators, and children alike (IBM Research, 2025). Additionally, the use of model cards and comprehensive documentation to enhance system transparency should be actively encouraged (Google, 2018). Google has issued a Responsible *AI Progress Report* which confirmed an ongoing publication of external model cards and detailed technical reports for its advanced AI models. These documents cover model creation, functionality, intended use, limitations, and performance metrics, serving as key transparency tools (Google, 2025).

5. Ensure Age-Appropriate and Inclusive Design

AI tools that specifically match children's cognitive, emotional, and social development stages should be designed and developed. Involve children in the design process to create inclusive technologies that effectively support learning and engagement for diverse users (Berson et al., 2025).

6. Support Multi-Stakeholder Approach

Educating all stakeholders, including parents, teachers, children, and others, is essential for raising awareness about digital rights and the risks associated with children's use of technology. Incorporating digital literacy and critical thinking into school curricula equips young people to engage with artificial intelligence (AI) and other digital tools in a responsible and informed way from an early age.

Collaboration among developers, policymakers, educators, child psychologists, and families is equally crucial. By working together, these groups can shape ethical frameworks that guide innovation. Multi-stakeholder collaboration fosters the creation of shared standards that promote child-centred and socially responsible technological advancement (Floridi & Cowls, 2019). Maintaining open and ongoing dialogue ensures that, as technology evolves, children's rights, safety, and well-being remain central to digital progress.

7. Amplify the Growth of Ethical AI Innovations in Child-Centred Technologies

Recognition and support of AI tools that already follow strong ethical standards is essential. These standards include practices that prioritise children's safety, well-being, autonomy, and rights throughout the

development and use of technology. Successful examples such as Scratch, Learning Passport, and Woebot for children show how AI can enhance learning and support mental health without compromising ethical principles. Greater emphasis should be placed on promoting and scaling these tools, as they serve as valuable models for responsible and child-centred innovation.

In conclusion, as AI and data-driven technologies become more integrated into children's lives, it is essential to ensure they are developed and used ethically, in children's best interests. By establishing strong ethical foundations, protecting data privacy, eliminating algorithmic bias, ensuring transparency, and designing age-appropriate and inclusive tools, we can create technologies that truly support children's rights and well-being. Education and collaboration among all stakeholders, developers, educators, families, and policymakers are essential to building a responsible digital future. By promoting ethical innovation and scaling proven child-centred tools, we can ensure that technology remains a force for good in the lives of all children.

CONCLUSION

The integration of Artificial Intelligence into technologies crafted for children marks a pivotal moment in their developmental journey, offering revolutionary pathways for personalised education, interactive play, and tailored support. Yet, as this comprehensive white paper has painstakingly detailed, realising the transformative benefits of AI responsibly is inextricably linked to our collective ability to rigorously address its profound ethical implications. The unique vulnerabilities of children, their evolving cognitive understanding, susceptibility to influence, and inherent right to privacy demand an ethical vigilance that transcends mere compliance. It is crucial to actively embed their fundamental rights and holistic well-being into the very fabric of AI design, deployment, and governance.

The imperative for robust ethical frameworks is paramount, ensuring that principles such as transparency, accountability, fairness, and privacy by design are not abstract ideals but actionable mandates guiding every stage of AI usage in the development of child-centred technologies. We must champion the creation of child-centred AI systems that are intuitively designed to be safe, equitable, and developmentally appropriate, preventing potential harms like data exploitation, algorithmic discrimination, or manipulative nudges. Furthermore, a dynamic and forward-looking regulatory landscape is crucial, capable of adapting to the rapid pace of technological change and enforcing stringent standards that protect children in increasingly complex digital environments.

Ultimately, securing a positive and empowering AI-driven future for children is not a task for any single entity but a shared, global responsibility. It calls for sustained, meaningful collaboration among all stakeholders: parents and guardians who guide their children's digital engagement; educators who impart critical digital literacy skills; technology developers who bear the primary responsibility for ethical innovation; policymakers who establish protective legal safeguards; and civil society organizations and researchers who advocate for children's rights and inform best practices.

By diligently integrating ethical AI literacy into educational curricula, promoting responsible and transparent technology design, strengthening legal mechanisms, and fostering continuous inter-sectoral partnerships, we can collectively construct a digital ecosystem where AI acts as a powerful catalyst for positive change, enriching children's lives without compromising their privacy, autonomy, or intrinsic value as individuals. This collective commitment will define whether AI becomes a cornerstone of equitable opportunity or a source of unforeseen challenges for the next generation.

REFERENCES

Achieng, R. (2023, March 29). AI & Children: Privacy, Trust, and Safety in the Digital Age. Strathmore University Centre for Intellectual Property and Information Technology Law. https://cipit.strathmore.edu/ai-children-privacy-trust-and-safety-in-the-digital-age/

Al-Zahrani, A. M. (2024). Unveiling the shadows: Beyond the hype of AI in Education. Heliyon, 10(9). https://doi.org/10.1016/j.heliyon.2024.e30696

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Berman, G., & Albright, K. (2017). Children and the data cycle: Rights and ethics in a big data world. UNICEF Office of Global Insight and Policy. https://defenddigitalme.org/wp-content/uploads/2018/02/IWP_2017_05_UNICEF_Ethicsbigdata.pdf

Berson, I. R., Berson, M. J., & Luo, W. (2025). Innovating responsibly: Ethical considerations for AI in early childhood education. AI Brain Child, 1(2). https://doi.org/10.1007/s44436-025-00003-5

Blanchard, E., Allard, E., & Roussel, N. (2021). Digital tools and children's learning: How interactive technologies support innovation and problem-solving. Journal of Educational Technology & Society, 24(3), 45–59.

Blanchard, E. G., Cox, A., Cooper, A., & D'Mello, S. (2021). Designing AI-driven educational technology for children: Ethical and pedagogical considerations. Computers & Education, 174, 104308. https://doi.org/10.1016/j.compedu.2021.104308

Blanchard, E., Razak, F., Harman, J., & Diaz, M. (2021). Children and artificial intelligence: Risks and opportunities (OECD Digital Economy Papers, No. 309). OECD Publishing. https://doi.org/10.1787/9b61f672-en

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Dafoe, A. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims (arXiv preprint arXiv:2004.07213). arXiv. https://doi.org/10.48550/arXiv.2004.07213

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

REFERENCES

Chubba, J. A., Missaoui, S., Concannon, S., Maloney, L., & Walker, J. A. (2021). Interactive storytelling for children: A case study of design and development considerations for ethical conversational AI. International Journal of Child–Computer Interaction, 32, Article 100403. https://doi.org/10.1016/j.ijcci.2021.100403

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv preprint arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Duolingo (2025). Duolingo ABC. https://abc.duolingo.com/

Eugenio, J. (2022). AI and mental health: Evaluating Replika's potential in adolescent well-being. AI & Society. https://doi.org/10.1007/s00146-022-01398-1

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behaviour therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomised controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Gándara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the black box: Detecting and mitigating algorithmic bias across racialised groups in college student-success prediction. AERA Open, 10, 23328584241258741. https://doi.org/10.1177/23328584241258741

Gasser, U., Maclay, C. M., & Palfrey, J. G. (2020). Child-centred AI: Designing AI for children's best interests. Berkman Klein Centre for Internet & Society at Harvard University. https://cyber.harvard.edu/publication/2020/child-centered-ai

Goodin, T. (2024, September 27). AI for children: Balancing innovation and ethics. EthicAI. https://ethicai.net/ai-for-children

Google. (2018). AI principles: Responsible AI practices. https://ai.google/principles/

Google. (n.d.). YouTube Kids parental guide. https://support.google.com/youtubekids/answer/6172308?hl=en

Google AI. (n.d.). Model cards. Retrieved June 2, 2025, from https://modelcards.withgoogle.com/

REFERENCES

Google. (2025). Introducing Gemini for kids under 13 with Family Link. Google Keyword Blog. https://blog.google/products/family/introducing-gemini-for-kids-under-13/

Google. (2025). Protecting teens and kids with age-appropriate experiences across Google. Google Safety Centre. https://safety.google/blog/protecting-teens-and-kids-across-google/

Google, (2025) Responsible AI Progress Report February 2025. P7. Retrieved from https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf?utm

Grant, J. (2024). Roundtable on emerging evidence on child online safety. The London School of Economics and Political Science. https://www.digital-futures-for-children.net/roundtable-child-online-safety

Hirsch, A. (2024, December 12). AI detectors: An ethical minefield. Centre for Innovative Teaching and Learning. https://citl.news.niu.edu/2024/12/12/ai-detectors-an-ethical-minefield/

Holloway, D., Green, L., & Livingstone, S. (2021). The changing landscape of childhood and technology: Designing for children's digital futures. Media, Culture & Society, 43(2), 243-259. https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1930&context=ecuworks2013

IBM Research. (2025). AI Explainability 360. Retrieved from https://research.ibm.com/blog/ai-explainability-360

Information Commissioner's Office. (2025). How does the right to erasure apply to children? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/children-and-the-uk-gdpr/how-does-the-right-to-erasure-apply-to-children/

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kannan, P. (2024, October 3). How harmful are AI's biases on diverse student populations? Stanford HAI. https://hai.stanford.edu/news/how-harmful-are-ais-biases-on-diverse-student-populations.

Kirkland, A., & Tsinaraki, C. (2025). Ethical considerations for the use of AI in early childhood education: A systematic review. Journal of Early Childhood Research, 23(1), 13–27. https://link.springer.com/article/10.1007/s44436-025-00003-5

Kidslox. (2024). The YouTube algorithm & kids: How to protect kids on YouTube. Retrieved from Kidslox.com. https://kidslox.com/guide-to/youtube-algorithm/

REFERENCES

Learning Passport. (2025). Transforming societies through education. https://www.learningpassport.org/about-learning-passport

Lieber, C. (2018, August 8). Tech companies use "persuasive design" to get us hooked. Psychologists say it's unethical. Vox. https://www.vox.com/2018/8/8/17664580/persuasive-technology-psychology

Lifelong Kindergarten Group at the MIT Media Lab. (n.d.). *Scratch: An overview.* Retrieved June 2025, from https://scratch.mit.edu/about

Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying online risk to children (CO:RE Short Report Series on Key Topics). Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE – Children Online: Research and Evidence. https://doi.org/10.21241/ssoar.71817

Mark. (2025, January 13). *7 best AI tools for parents in 2025* [Blog post]. *Canopy.* https://canopy.us/blog/ai-tools-for-parents/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Miller, S., Saidy, E., & Solomon, H. (2024, February 20). Harnessing artificial intelligence for child protection: An ethical roadmap. Save the Children International. https://www.savethechildren.net/blog/harnessing-artificial-intelligence-child-protection-ethical-roadmap

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

MIT Media Lab. (2025). Scratch: An overview. https://scratch.mit.edu/about

PwC Nigeria. (2025). AI in Nigeria: Opportunities, challenges and strategic pathways. PwC Nigeria, Lagos Business School, & Microsoft. https://www.pwc.com/ng/en/assets/pdf/ai-in-nigeria-%20.pdf

Ragone, G., Buono, P., & Lanzilotti, R. (2024, April 22). Designing safe and engaging AI experiences for children: Towards the definition of best practices in UI/UX design [Workshop paper]. Workshop on Child-centred AI Design, CHI 2024. arXiv. https://arxiv.org/abs/2404.14218

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., & Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y

REFERENCES

Radesky, J. S., Bridgewater, E., Black, S., O'Neil, A., Sun, Y., Schaller, A., Weeks, H. M., & Campbell, S. W. (2024). Algorithmic content recommendations on a video-sharing platform used by children. JAMA Network Open, 7(5), e2413855. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/10.1001/jamanetworkopen.2024.13855

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., & Kafai, Y. (2009). Scratch: Programming for all. Communications of the ACM, 52(11), 60–67. https://doi.org/10.1145/1592761.1592779

Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson. (ISBN 9780134610993)

Safer Schools (2022, January 20). *What You Need To Know About. . .Replika.* https://oursaferschools.co.uk/2022/01/20/replika-ai-friend/

Save the Children International. (2024). Harnessing artificial intelligence for child protection: An ethical roadmap. Retrieved from https://www.savethechildren.net/blog/harnessing-artificial-intelligence-child-protection-ethical-roadmap.

Simas Velez. (2023). The impact of AI-powered devices on children's privacy and surveillance: Ethical and responsible use of technology. SimasVelez.com. https://www.simasvelez.com/the-impact-of-ai-powered-devices-on-childrens-privacy-and-surveillance-ethical-and-responsible-use-of-technology

The Adventures of Muna. https://thesafetychic.com/adventures-of-muna/

UNICEF. (2020). UNICEF and Microsoft launch global learning platform to help address COVID-19 education crisis. https://www.unicef.org/press-releases/unicef-and-microsoft-launch-global-learning-platform-help-address-covid-19-education

UNICEF, Office of Global Insight and Policy. (2021, November). Policy guidance on AI for children 2.0. UNICEF. https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf

UNICEF. (2023). Giga: Connecting every school to the internet. https://giga.global

Vesselinov, R., & Grego, J. (2012, December). Duolingo effectiveness study. City University of New York. https://theowlapp.health/wp-content/uploads/2022/04/DuolingoReport_Final-1.pdf

REFERENCES

UNICEF. (2024). Nigeria Learning Passport. https://www.unicef.org/nigeria/press-releases/nigeria-learning-passport-reaches-one-million-subscribers-milestone-educational

University College London. (2021). Fair-AIEd: Participatory design approaches to creating ethical AI for international education and development [Funded project summary]. UK Research and Innovation. https://gtr.ukri.org/projects?ref=MR%2FT022493%2F1

Vesselinov, R., & Grego, J. (2012, December). Duolingo Effectiveness Study [Research report]. City University of New York & University of South Carolina. Retrieved from https://theowlapp.health/wp-content/uploads/2022/04/DuolingoReport_Final-1.pdf

Woebot Health. (2024). Woebot for adolescents (Woebot Adolescent 1.0) Non-prescription digital mental health software. https://woebothealth.com/img/2024/07/Woebot-for-Adolescents-User-Instructions-for-Use_July-18th-2024.pdf

Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024, June 15). Exploring Parent-Child perceptions on safety in Generative AI: Concerns, mitigation strategies, and design implications. arXiv.org. https://arxiv.org/abs/2406.10461

THE Safety Chic

www.thesafetychic.com