OPPORTUNITIES AND RISKS OF GENAI CHATBOTS FOR CHILDREN

Thao Do-Ngoc

December 2024

Acknowledgement: This research was supported by the INCLUDE+/Ofcom Fellowship. A special thanks to Andreas Haggman, Valery Otieno, Victoria Jaynes, Halla Bjork Kristijansdottir, Victoria Taylor and Martha Kirby from Ofcom, and Prof. Emma Carmel from the University of Bath for your valuable feedback. Thanks to Niamh Cashell, Rosemary Wilkinson, and Helen Thornham for your great support and facilitation.

Thao-Do Ngoc is current undertaking Integrated PhD in Accountable, Responsible, and Transparent AI at the ART-AI center, University of Bath. Her research focuses on the responsible use of AI in addressing human trafficking. https://orcid.org/0000-0002-0015-892X

Disclaimer: This paper represents the views and opinions of the author and should not be taken as a statement of Ofcom policy/opinion.

Table of Contents

OVERVIEW

The rapid advancements in Generative AI (GenAI) technology have significantly changed how people interact with digital tools, particularly conversational AI such as chatbots. Leveraging sophisticated machine learning and natural language processing capabilities, these AI-driven chatbots (hereby called *GenAI chatbots*) can generate human-like responses, making them increasingly integrated into various sectors such as education, business and entertainment. The potential benefits of GenAI chatbots for children are significant, ranging from personalised educational support to enhanced creativity and social interaction.

However, the rapid adoption of GenAI chatbots, especially among children, also raises significant concerns about safety, privacy, and developmental impacts. Children are defined as individuals under the age of 18, according to UN Conventions on the Rights of the Child[1], covering the developmental range from early childhood to adolescence (0-18 years). Children represent a unique and vulnerable user group due to their ongoing cognitive, social, and emotional development. This makes them more susceptible to risks such as exposure to harmful content, misinformation, emotional dependency, and biased outputs. Understanding how children interact with GenAI requires acknowledging that that risks, harms, and opportunities manifest across different stages of development, age groups, cultural contexts, and socioeconomic backgrounds. Although the legal definition of a child ends at age 18, it is important to consider that brain development, particularly the prefrontal cortex, continues until around age 25. This ongoing development suggests that protective measures should not cease at age 18 but acknowledge that. The introduction of AI into children's ecosystems, including family, school, and social interactions, poses a novel and complex variable that can shape developmental trajectories, with potential consequences for cognitive and socio-emotional growth.

While much research has focused on the potential benefits, risks and harms associated with GenAI in general applications, including AI safety[2], bias and fairness in chatbots[3], and the benefits and risks of personalised language models[4], there is a lack of comprehensive studies examining its implications for children. This research gap is particularly concerning given that children are especially vulnerable users. This report seeks to bridge that gap by providing a comprehensive analysis of how GenAI tools can support children's learning and development, while also identifying potential risks and harms, and addressing safety concerns. It seeks to inform stakeholders, including policymakers, educators, developers, parents, and the broader public, providing critical evidence to inform further research and guide regulatory actions. With respect to "for children", the study focuses on: GenAI chatbots that were *explicitly designed for* children (but not necessarily have to be used by children), systems that children *interact with* (not specifically designed for children but could be accessed by them), and most broadly, systems that may *impact* children[5].

[1] UN. (1989). The United Nations Convention on the Rights of the Child.
[2] Chua, J., Li, Y., Yang, S., Wang, C., & Yao, L. (2024). AI Safety in Generative AI Large Language Models: A Survey.
[3] Xue, J., Wang, Y. C., Wei, C., Liu, X., Woo, J., & Kuo, C. C. J. (2023). Bias and fairness in chatbots: An overview.
[4] Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback.
[5] UNICEF. (2021). Policy guidance on AI for children.

This study employs a scoping review to comprehensively map the existing literature on the use and impact of GenAI chatbots for children. It draws from a wide range of disciplines, including GenAI technical development, child psychology, AI ethics, education, social sciences, human-computer interaction (HCI), and laws and regulatory frameworks to provide a holistic understanding of the current landscape and emerging issues. A scoping review is particularly suitable for this purpose as it facilitates the exploration of broad and emerging topics where evidence is diverse and evolving[6]. This approach helps identify research gaps, clarify key concepts, and guide future research and policy development by integrating finding types of evidence, including grey literature and policy documents[7].

However, it is important to acknowledge several limitations of the report. The limited literature on GenAI's impact on children and the absence of longitudinal studies limits the understanding of the long-term developmental effects of GenAI on children. While the report aims to address diverse child experiences, it may not fully account for all developmental stages, cultural contexts, or socioeconomic backgrounds. Additionally, there may be a publication bias toward studies that emphasise the benefits or risks of AI. The rapid evolution of GenAI technology could quickly render some findings outdated as new risks or benefits emerge.

This report addresses the unique intersection of GenAI technology and child development. It contributes to the discourse on AI ethics and safety by integrating insights from various disciplines, including child psychology, AI technical development, human-computer interaction, and regulatory frameworks. This interdisciplinary approach allows for a comprehensive analysis that not only identifies gaps in existing research but also offers a holistic understanding of the complexities involved in children's interactions with GenAI. The report can also serve as a practical resource for stakeholders—policymakers, educators, developers, and parents—by providing evidence-based recommendations for safer AI design and implementation. It calls for child-focused regulations that account for the diversity of children's developmental stages and social contexts, advocating for protective measures that can mitigate risks while maximising the benefits of AI. As GenAI technologies continues to evolve, this report seeks to shape ongoing discussions about safe and responsible AI use for children, balancing approach that supports learning, creativity, and safety for children.

[6] Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology*, *8*(1), 19-32.

[7] Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18*, 1-7.

EXECUTIVE SUMMARY

Perceptions and current use of children of GenAI tools (UK context)

Children's perceptions of GenAI chatbots vary widely and influence their interactions with the technology. Younger children often see AI as friendly and attribute both human and non-human qualities, while older children view AI as intelligent but not human-like. However, children generally overestimate the abilities of GenAI, seeing it as smarter than themselves or similar to search engines or advanced assistants such as Alexa. Misunderstandings about GAI's limitations can lead to overtrust and vulnerability to misinformation. Educational activities play a crucial role in addressing these misconceptions by highlighting GenAI's limitations, enhancing a more critical and informed perspective

Awareness and usage of GenAI tools among UK children have surged in 2024 compared to previous years. Younger children (under 13) are increasingly adopting these tools despite age restrictions, with similar usage rates across socio-economic groups. However, differences in usage among ethnicities need further evidence, and preferences of GenAI tools may vary by gender. Children use GenAI for various purposes: including education, entertainment, socialisation, and experimentation, but often lack critical engagement. Children's use of GAI tools is often beyond parents' and teacher's understanding and knowledge. Effective parental controls are limited, highlighting the need for safer, child-centered design

Opportunities of GenAI chatbots for children

GenAI chatbots offer several benefits for children, including personalised educational support, creative exploration, and inclusivity for diverse learning needs. Key opportunities include:

- **Educational enhancement**: GenAI chatbots can provide personalised learning experiences tailored to individual needs, helping children with homework, language learning, and concept reinforcement. They can support differentiated instruction and serve as valuable resources in both formal and informal educational settings.
- **Creative expression and exploration**: These chatbots enable children to engage in creative tasks, such as writing stories, generating artwork, or coding, thereby enhancing curiosity and learning through play. Tools such as text-to-image and text-to-text applications are popular among children for experimenting with ideas and boosting creativity.
- **Support for special needs**: For children with social communication challenges, such as those on the autism spectrum, GenAI chatbots can simulate social scenarios to improve communication skills. They offer a safe space for practicing social interactions, recognising emotions, and enhancing language skills, and helping to promote a more inclusive learning environment.
- **Online safety and abuse prevention**: GenAI chatbots can be utilised to raise awareness about online threats, detect harmful content, and flag inappropriate interactions. They can provide

educational support on recognising online dangers and responding appropriately to potential abuse, thereby contributing to child protection efforts.

Risks and harms of GenAI chatbots for children:

Despite their potential, the use of GenAI chatbots by children is not without risks. Key concerns include:

- **Exposure to harmful and inappropriate content**: GenAI chatbots may generate age-inappropriate, offensive, or harmful content due to training on large, unchecked datasets. Even with safety measures in place, chatbots can produce toxic responses or misinformation, which can negatively impact children's emotional, psychological well-being, and normal development.
- **Emotional and psychological risks**: The use of GenAI chatbots as digital companions may lead to emotional dependency and hinder the development of social skills. Children might anthropomorphise chatbots, forming unhealthy attachments that affect their ability to build meaningful human relationships. Additionally, these tools may not adequately detect or respond to distress signals, potentially exacerbating mental health issues.
- **Misinformation and cognitive risks**: Children may overtrust chatbot-generated content, equating GenAI with factual accuracy. The inability of chatbots to consistently verify information can result in the spread of misinformation, influencing children's decision-making and worldview. This risk is heightened by the difficulty children face in distinguishing AI-generated responses from fact-checked information.
- **Privacy and data security**: The collection and handling of children's data by GenAI tools raise privacy concerns, with risks of data breaches, identity theft, and misuse of personal information. Insufficient safeguards for protecting sensitive data exacerbate these risks.

Gaps/challenges in regulating GenAI for children

Regulatory challenges:

- Principles-based AI regulation is flexible but can lead to misinterpretation, such as misuse of the "best interests" standard. Guidelines often lack actionable advice for safeguarding children's rights.
- Complexity of ethical AI for children: Existing frameworks overlook developmental stages, guardians' roles, and human-centered evaluations, focusing mainly on technical aspects.
- Limited child participation: Despite interest, meaningful child involvement in AI policy remains scarce due to a lack of resources and expertise.

Design/technical challenges:

- The shift from designing with children (education, safety) to a holistic, child-centered approach is incomplete. The need for comprehensive AI design principles remains unmet.
- Evaluating GenAI chatbots is challenging as current metrics fail to capture conversational quality, and model transparency is limited.

Safety measures challenges:

- Age verification and parental controls: Current safety tools present privacy concerns and potential discrimination, while over-reliance on parental controls can undermine children's autonomy. More research is needed for inclusive and balanced approaches.

Collaboration challenges:

- Lack of collaboration between AI, law, and psychology hinders the development of child-centric principles, calling for more integrated approaches.

Recommendations

To ensure the safe and effective development and deployment of GenAI chatbots for children, a multi-layered approach is needed, addressing the responsibilities of developers, regulators, and other stakeholders.

1. **For GenAI developers and deployers:**
    - **Transparency obligations:** Developers should disclose training data sources, performance metrics and harmful content incidents. Continuous monitoring and transparent reporting are needed throughout the model lifecycle. Information should be accessible to all, especially regarding AI's impact on children, using AI labeling systems and clear disclosures.
    - **Independent auditing & bias mitigation:** Proactive audits of training data should address bias, with synthetic data used to counteract biases in online sources. Controlled releases of models balance public scrutiny with misuse risks, while child safety data should be accessible to researchers.
    - **Child-Centered approach:** Involves children and stakeholders in AI design to ensure developmentally appropriate tools. Goes beyond superficial features to prioritise children's autonomy, needs, and perspectives, empowering them in the design process.

2. **For regulators:**
    - **Incorporating children's voices**: Actively involve children's perspectives in regulatory decisions, using research on their online experiences to inform policies and improve AI design.
    - **Providing practical guidelines**: Establish clear expectations for implementing ethical AI principles, including a "child online safety tracking database" to document safety updates and make them publicly accessible.
    - **Establishing standards and collaboration**: Promote a bottom-up approach that brings together developers, child protection specialists, and policymakers to develop best practices. Collaborate internationally to set unified standards for online safety.
    - **Using regulatory experimentation tools**: Employ tools like regulatory sandboxes to test new safety features in controlled environments, allowing for iterative improvements in AI regulation.
    - **Quantifying potential harms**: Develop metrics to assess online risks for children based on factors such as age, severity, scale, and existing harm reduction measures, guiding regulatory priorities

3. **For Ecosystems (Parents, Teachers, Caregivers, Civil Society, and the Public):**
    - **Decentralised content moderation:** Users, including "trusted flaggers" (NGOs, volunteers), can flag harmful GenAI content. Flaggers report issues to developers, and companies prioritize addressing flagged content to prevent misuse. A mix of centralised and decentralised monitoring enhances management of harmful content.
    - **Civil society involvement:** Civil society groups can file "super-complaints" and advocate for faster responses from platforms. Regulators should collaborate with civil society for evidence and harm context.
    - **Parent-AI collaborative system:** Dynamic content moderation allows parents to adapt filters for their child's needs. Involving children in co-design ensures solutions are more effective and relevant.

4. **For children:**

- **Child participation in AI development:** Children can be users, contributors, or innovators in shaping AI policies and practices.
- **AI literacy and ethics education:** Educational activities could demystify AI, teaching children about its limitations, risks, and ethical considerations. Age-appropriate ethics education enhance critical thinking and helps children avoid over-reliance on AI.
- **Youth-shaped algorithms and platform moderation:** Children can help design algorithms and report harmful content, promoting positive experiences. A child-friendly feedback system should allow reporting of issues directly to moderators. Platform moderation needs to be inclusive, age-sensitive, and diversity-aware.

Despite the advanced capabilities of GenAI chatbots, they also present significant drawbacks, especially when considering their societal and ethical implications for children as users. While much research has focused on the potential risks and harms associated with GenAI, including AI safety[8], bias and fairness in chatbots[9], and the benefits and risks of personalised language models[10], there has been limited exploration of the specific risks and harms these technologies pose to children. This research gap is particularly concerning given that children are especially vulnerable users. Children's unique perspectives and limited life experiences shape how they engage with AI and interpret its outputs. Due to their ongoing emotional, cognitive, and social development, they are more susceptible to risks such as manipulation, bias, and data privacy breaches[11]. Additionally, children are still forming their identities and are more influenced by social comparisons and peer validation, which GenAI can exacerbate. AI systems, by amplifying negative biases, can have long-lasting effects on children's self-esteem, mental health, and identity development[12]. Moreover, with developing impulse control and emotional regulation, children may lack the autonomy and decision-making capacity necessary to navigate these technologies safely. This scoping review aims to highlight both the opportunities and risks posed by GenAI for children, providing critical evidence to inform further research and guide regulatory actions.

I. BACKGROUND

1.1. Understanding GenAI chatbots

Conversational Artificial Intelligence (AI) refers to AI-driven technologies of chatbots, intelligent personal assistants (IPAs), robots, and multimedia interfaces[13]. Chatbots are software applications

[8] Chua, J., Li, Y., Yang, S., Wang, C., & Yao, L. (2024). AI Safety in Generative AI Large Language Models: A Survey.
[9] Xue, J., Wang, Y. C., Wei, C., Liu, X., Woo, J., & Kuo, C. C. J. (2023). Bias and fairness in chatbots: An overview.
[10] Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback.
[11] Reich, S. M., Starks, A., & Su, Z. (2024). One (Adult) Size Does Not Fit All: The Importance of Development in Digital Design and Utilization. *Youth Wellbeing in a Technology Rich World*.
[12] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.
[13] Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*(1), 48-63.

designed to simulate human conversation and interact with users through various modalities, including text, voice, graphical interfaces, and potentially other emerging technologies[14][15]. Chatbots can be either custom-built for research purposes or off-the-shelf applications such as ChatbotGPT, Cleverbot and Replika. Chatbots can function as standalone products or be integrated into software, such as online platforms, websites, digital assistants, mobile apps or messaging services[16]. IPAs include systems such as Amazon Alexa and Google Assistant, which are widely used for voice-activated tasks such as setting reminders, controlling smart devices, answering queries, and providing real-time information. Multimedia platforms such as virtual reality, extended reality, or embodied conversational agents are considered part of conversational AI when they are enhanced with AI capabilities. This report concerns chatbots with a focus on GenAI chatbots. With respect to "for children", the study focuses on: GenAI chatbots that were *explicitly designed for* children (but not necessarily have to be used by children), systems that children *interact with* (not specifically designed for children but could be accessed by them), and most broadly, systems that may *impact* children[17].

The history of chatbots has evolved from simple rule-based systems to advanced AI-driven models[18]. (1) Early foundations (1950s-1970s): The concept began with Alan Turing's 1950 "Turing Test." Early chatbots like ELIZA (1966) and PARRY (1972) used pattern matching to simulate conversations but lacked true understanding. **(2) Advancements with Artificial Intelligence Markup Language (AIML) (1990s):** Richard Wallace's ALICE (1995) used AIML to manage conversations, improving on earlier approaches but still relying on scripted responses. **(3) Machine Learning Integration (2000s):** Chatbots began using Natural language processing (NLP) and machine learning, with examples such as SmarterChild demonstrating real-time interaction capabilities. **(4) Modern AI Chatbots (2010s-present):** Cognitive assistants such as Siri, Alexa, and Google Assistant emerged, marked a shift towards chatbots that could process and respond to complex queries. Transformer-based models, such as GPT-3 and Blenderbot using deep learning, enabled chatbots to understand context and generate coherent responses.

Compared to traditional chatbots, GenAI chatbots are more advanced in their ability to handle language and understand context. While traditional chatbots rely on predefined rules and scripted responses, limiting them to simple, structured tasks, GenAI chatbots use deep learning and natural language processing (NLP). They are trained on vast datasets, enabling them to engage in more dynamic, natural, and contextually relevant conversations. They are equipped with problem-solving abilities, enabling them to manage complex queries and perform creative tasks such as content generation. They are highly scalable, integrate seamlessly with other systems, and can access real-time data and external APIs. . This evolution has been driven by advancements in computing power, NLP techniques, and the availability of large datasets[19]. However, generative AI chatbots can require significant computational power and large datasets

[14] IBM. What is a chatbot? Available at: https://www.ibm.com/topics/chatbots (accessed 22nd Oct 2024)
[15] Hasal, M., Nowaková, J., Ahmed Saghair, K., Abdulla, H., Snášel, V., & Ogiela, L. (2021). Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation: Practice and Experience*, *33*(19), e6426.
[16] Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924.
[17] UNICEF. (2021). Policy guidance on AI for children.
[18] Al-Amin, M., Ali, M. S., Salam, A., Khan, A., Ali, A., Ullah, A., ... & Chowdhury, S. K. (2024). History of generative Artificial Intelligence (AI) chatbots: past, present, and future development.
[19] Al-Amin, M., Ali, M. S., Salam, A., Khan, A., Ali, A., Ullah, A., ... & Chowdhury, S. K. (2024). History of generative Artificial Intelligence (AI) chatbots: past, present, and future development.

for training, which can be costly. Their responses can also be unpredictable, sometimes producing inaccurate, irrelevant, or inappropriate outputs. Moreover, there are ethical concerns regarding the potential for biased, harmful, or misleading content generation, requiring careful oversight to mitigate risks.

Table 1: Taxonomy of chatbots

Khennouche et al. (2024)[20] presented a taxonomy of chatbots, based on the models used: rule-based, retrieval-based, generative, and hybrid.

Rule-based chatbots operate using predefined rules, such as decision trees or flowcharts, to deliver pattern-based responses. They typically start with greetings and guide users through Yes/No questions or selectable options to shape the conversation. This type includes decision tree chatbots, which follow a step-by-step process (e.g., welcome messages and user selections), and keyword chatbots, which respond based on specific user-input keywords. Although these systems lack flexibility, they provide consistent and predictable responses, making them suitable for straightforward use cases.

Retrieval-based chatbots use a predefined database of questions and answers to match user input with existing responses. Unlike generative chatbots, they do not create new responses but rely on stored answers. Developing these chatbots involves building a dataset of potential queries and corresponding replies. They work well for straightforward and consistent questions, such as in customer support, where prompt and reliable answers are needed. However, they may struggle with complex queries due to their limited response options.

The difference between rule-based and retrieval-based chatbots lies in their response generation methods. Rule-based chatbots generate replies by following predefined rules and decision paths, while retrieval-based chatbots choose responses by matching user input to a set of existing responses stored in a database.

Generative chatbots are capable of providing more interactive and personalised conversations by generating responses from scratch, allowing them to adapt to new contexts and situations. Because they create responses autonomously, there is a risk of producing irrelevant or inappropriate replies, making thorough testing and monitoring essential to ensure quality and safety. Generative chatbots typically use neural network models and large datasets for training, though domain-specific datasets can be sufficient for specialised applications, such as FAQs.

Hybrid chatbots combine retrieval-based and generative methods. They start by matching user queries to a dataset of predefined question-answer pairs using techniques such as keyword matching or vector similarity. When a suitable answer is not found, the chatbot resorts to a generative model to create a response, allowing it to address more complex or unique queries beyond the FAQ database. For example, a chatbot integrating IBM Watson with Moodle helped students in higher education by improving the accuracy of responses to course-related questions. Another example involved a customer support chatbot that used DialogFlow for language understanding and additional modules to enhance response selection. Additionally, MILABOT, a hybrid chatbot for the Amazon Alexa Prize, combined multiple AI techniques and used reinforcement learning to improve its conversational abilities.

[20] Khennouche, F., Elmir, Y., Himeur, Y., Djebari, N., & Amira, A. (2024). Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs. *Expert Systems with Applications*, *246*, 123224. The study is based on a review of 107 papers on chatbots from 2013 to 2022 (with a focus on FAQ chatbots).

Bandi et al., (2023)[21] further classified GenAI models based on architectural characteristics: variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion models, transformers, language models, normalising flow models, and hybrid models. The author also includes **classification of input and output formats**: text to text, text to image, text to audio/speech, text to code and code to text, image to text, and visual content generation (image to image, text to video, text+video to video, video to video, image+text to image), text-driven 3D content generation (text to 3D image, text to 3D animation), text to molecule or molecule structure, tabular data to tabular data, text to knowledge graph and knowledge graph to text, and road network to road network.

Core components of GenAI chatbots[22]: **(1) Foundation Model:** The foundation model is the large-scale language model that serves as the core of a GenAI chatbot. These models are trained on diverse datasets, such as books, articles, and online content, to learn the patterns of human language, including grammar, syntax, and semantics. The training process enables the model to understand prompts and generate relevant responses. Models such as GPT-3, GPT-4, and other transformers have millions or even billions of parameters, allowing them to generate detailed and nuanced text. **(2) User Interface (UI):** The UI is the platform through which users interact with the chatbot. It can be text-based (e.g., web chat windows, messaging apps), voice-based (e.g., smart speakers), or multimodal interfaces that combine various interaction methods, such as text, voice, and images. **(3) Prompting and Input Processing:** The way prompts are presented significantly impacts the chatbot's response quality. Prompt engineering involves structuring input in a manner that guides the model toward generating useful output. For example, a well-phrased prompt can specify the style, tone, or format of the desired response. The ability to fine-tune prompts ensures that responses meet user expectations and align with specific tasks. **(4) Output generation:** The output generation process involves the model producing a response based on the input prompt and its underlying architecture. The quality of the output is influenced by factors such as the model's training data, prompt clarity, and any applied fine-tuning for specialised tasks.

Developing GenAI chatbots involves distinct steps and considerations[23]: **(1) Creation**: The process starts with defining goals, understanding user needs, and selecting large datasets for training. Development uses AI frameworks such as TensorFlow or PyTorch and language models such as GPT. The chatbot is tested and deployed on cloud platforms to handle high computational demands, with integration via APIs for multi-channel interactions. **(2) Training**: GenAI chatbots train on vast text corpora to generate human-like responses. Fine-tuning with domain-specific data enhances relevance, but training requires significant computational resources, specialised hardware, and ongoing updates. **(3) Integration**: GenAI chatbots are integrated with messaging apps, voice assistants, or websites, providing dynamic responses across different interfaces. Website deployment allows more customisation, while messaging platforms enhance accessibility. (4) **Conversational abilities**: Chatbots can adopt personas (e.g., famous figures) to enhance user engagement. They can understand context, maintain multi-turn conversations, and offer coherent responses across diverse topics. Managing response quality is key to avoiding irrelevant or inappropriate

[21] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260. The study is based on a review of 122 papers on GenAI published from 2014 to 2023 (90% of the papers were published from 2017 till present).

[22] Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with applications*, *2*, 100006.

[23] Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with applications*, *2*, 100006.

outputs. **(5) Other considerations**: Developers must address ethical concerns such as bias, ensure data privacy, manage latency, and optimise computational costs. Supporting multiple languages often requires advanced NLP techniques and model fine-tuning.

However, in the development of chatbots, it is not always necessary to build models from scratch. Many modern chatbots, such as Snapchat's My AI, leverage existing large language models such as OpenAI's GPT technology. This approach allows developers to integrate advanced conversational capabilities without starting from the ground up. By **fine-tuning pre-existing models or utilising APIs**, chatbots can be tailored to specific use cases, enabling quick deployment and reducing development cost.

Valz (2023)[24] highlights how safety risks can occur from the design and deployment of GenAI chatbots in different stages: (1) Pre-training: Models are trained on large datasets to learn language patterns, but this phase can introduce toxic content, as filtering is rare for foundational models, posing safety risks. **(2) Fine-tuning:** Domain-specific fine-tuning refines task relevance but cannot fully eliminate toxic content, as models retain associations from pre-training; **(3) Alignment**: Techniques such as reinforcement learning and classifiers help prioritise safe, constructive responses by guiding the model's behaviour toward desired outcomes**; (4) Post-deployment controls**: Despite thorough training, harmful outputs may still occur. Post-deployment measures include limiting interactions, using additional filters, and scripted responses, though bypassing these safeguards is still possible.

Key limitations and challenges of chatbots that impact their performance and raise societal and ethical concerns[25][26] include: (1) **Hallucinations**, where chatbots generate inaccurate or misleading answers due to unpredictability; (2) **Bias**, as chatbots may inherit prejudices present in their training data, leading to offensive or unfair outputs; (3) **Privacy and security risks**, since chatbots require large amounts of data, creating potential vulnerabilities if data is not adequately protected; (4) **Black-box nature and interpretability**, with opaque decision-making processes that make it difficult to trust their outputs or diagnose problems. Their effectiveness heavily depends on the quality of training data, which can be a limitation in niche applications where data is scarce; (5) **Environmental concerns**, as the computational demands of GAI models require significant power, raising sustainability issues; (6) **Context and semantic understanding**, where chatbots struggle with accurately determining the context of a conversation or interpreting words with multiple meanings, leading to potential misinterpretations; and (7) **Grammar and language structure**, where chatbots face challenges with ambiguous questions, grammatical errors, or handling multiple languages with different rules

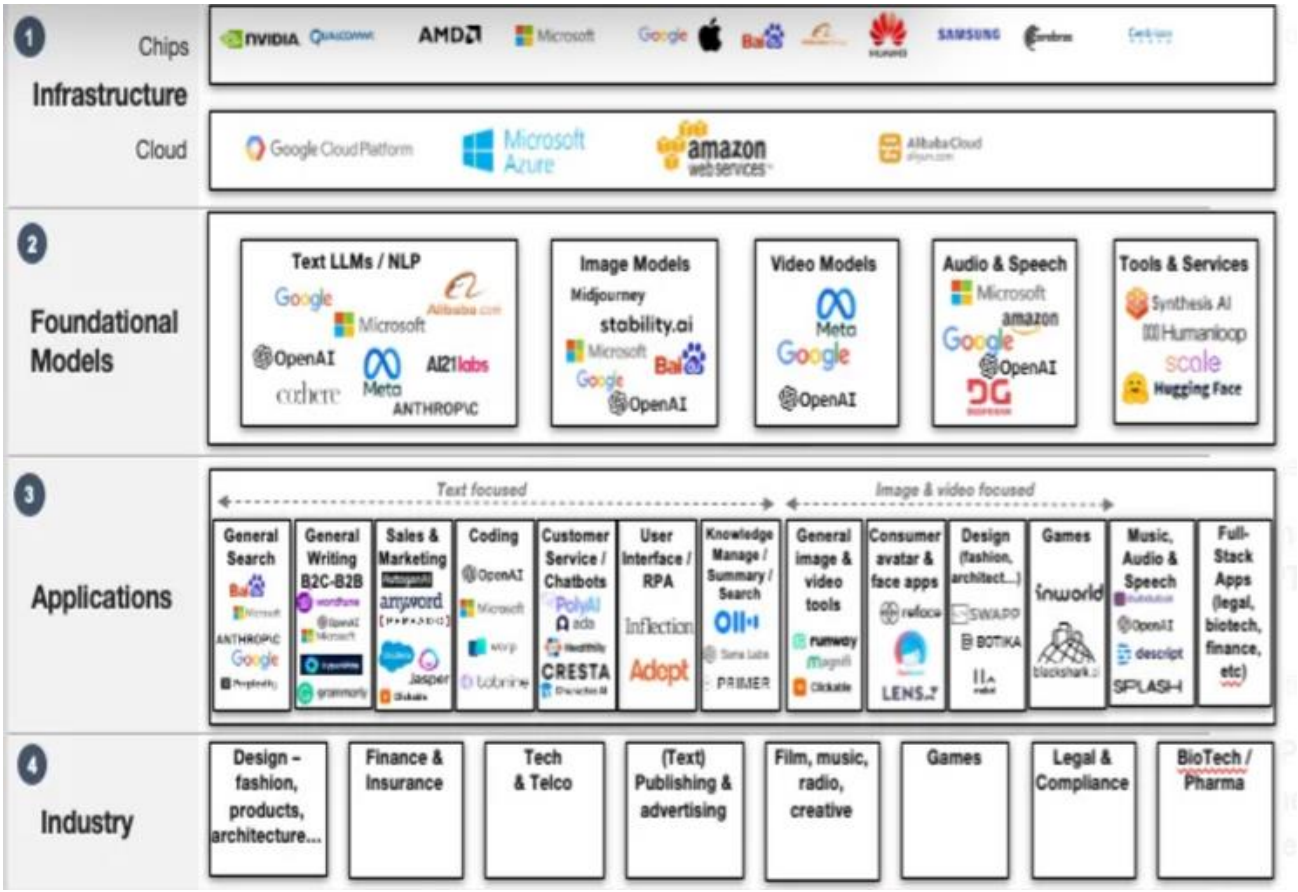
[24] Valz, D. (2023). Personalisation: Why the relational modes between Generative AI chatbots and human users are critical factors for product design and safety.

[25] Khennouche, F., Elmir, Y., Himeur, Y., Djebari, N., & Amira, A. (2024). Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs. *Expert Systems with Applications*, *246*, 123224. The study is based on a review of 107 papers on chatbots from 2013 to 2022 (with a focus on FAQ chatbots).

[26] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

The GenAI ecosystems consists of multi layers[27][28]: (1) Infrastructure providers: Companies that supply GPUs, hardware, and cloud-computing services to run GenAI software. **(2) Foundational Models:** Large language models (LLMs) created by companies such as OpenAI, Google, and Meta, which serve as the basis for building AI applications. **(3) Development tools:** SDKs, frameworks, and libraries that help in training models and building applications. **(4) Applications:** A rapidly growing number of AI-powered applications targeting both general users ("horizontal vendors") and specific industries ("vertical orders"). Established companies are also integrating GenAI into their existing products. **(5) Data Providers:** Entities that supply data used for training GenAI models. **(6) Regulatory Infrastructure:** Emerging rules and regulations that governments will implement to manage the use and development of GenAI.

Figure 2: GenAI ecosystems



Source: Greenman, 2023[29]

[27] Greenman, S. (2023). Who will make money from the generative AI gold rush (accessed 20th Oct 2024)
[28] Cusumano, M. A., Farias, V. F., & Ramakrishnan, R. (2024). Generative AI as a New Platform for Applications Development. *An MIT Exploration of Generative AI*
[29] Greenman, S. (2023). Who will make money from the generative AI gold rush (accessed 20th Oct 2024)

1.2. Perceptions of children towards GenAI chatbots

Children of different ages have varying perceptions of GenAI chatbots, which has important implications for assessing the risks associated with GenAI exposure across age groups and designing age-appropriate safety measures. Additional factors such as prior exposure to technologies, the surrounding environment (e.g., families or schools)[30], personal traits[31], educational background and cultural influences[32] can shape and influence how children perceive and interact with the chatbots. Previous literature on children's interactions with rule-based chatbots (e.g., Alexa, Siri) can help shed light on the interactions between children and GenAI chatbots while also acknowledging the differences between the two types.[33] Recent studies on advanced AI assistants (Intelligent Personal Assistants (IPA)) indicate that children do not distinguish between humans and AI as strictly as adults do, though this perception shifts with age and technological experience[34]. Young children **(3-6 years old)** often perceive those AI assistants as friendly and truthful[35]. They tend to view chatbots as something between human and machine, sometimes attributing both human-like qualities (e.g., emotions) and inanimate characteristics to them. This reflects their ongoing construction of understanding when it comes to these technologies[36].

Older children **(7-10 years old)** start interacting more frequently with AI assistants and often consider these agents to be more intelligent than themselves[37]. A study conducted with primary-school children **aged 5-12** in Scotland also found similar results[38]. This tendency is crucial to understand as it influences their trust in these systems. While children in this age groups may associate AI with positive attributes and express excitement about their use, they do not generally view AI as human-like or believe it possesses human capabilities, such as experiencing emotions like anger or upset. However, they do believe it is wrong to be rude to conversational assistants. Despite these positive associations, children lacked a proper understanding of data privacy and security issues. Another study of children **aged 4-11** found that younger children attributed more agent-like features (e.g., having experiences, minds, and deserving moral

[30] Garg, R., & Sengupta, S. (2020). He is just like me: a study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), 1-24.
[31] Grassini, S., & Koivisto, M. (2024). Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. *Scientific Reports*, *14*(1), 4113.
[32] Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024, May). How Culture Shapes What People Want From AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
[33] Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE access*, *9*, 110558-110572.
[34] Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced ai assistants.
[35] In this study, 26 participants (3-10 years old) interact with Amazon Alexa, Google Home, Cozmo, and Julie Chatbot. See more: Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). " Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction. In *Proceedings of the 2017 conference on interaction design and children* (pp. 595-600).
[36] Studies of 28 children aged 3-6 years old. See more: Xu, Y., & Warschauer, M. (2020). What are you talking to?: Understanding children's perceptions of conversational agents. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-13).
[37] Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced ai assistants.
[38] Andries, V., & Robertson, J. (2023). Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence*, *5*, 100176.

treatment) to technologies[39]. However, these attributions declined with age and differed based on the type of technology (e.g., Alexa, Roomba, and Nao), indicating that different designs affect how children perceive and interact with these technologies.

Children can have misconceptions about GenAI chatbots, overestimate their abilities, or overtrust in the systems. Secondary school (key stage 3) students were found to equate AI with automation and robotics[40]. Secondary school (key stage 4) students also tend to perceive GenAI as a search engine or a vast database[41]. Generally, children struggle to differentiate GenAI from other forms of conventional AI (e.g., Google search, Alexa, Siri). They described GenAI as "*like an advanced Alexa*", or "*similar to Google or other [search engines], which are often right these days*." However, children can also display a critical understanding of GenAI mechanisms, such as one child noting "*Depending on when you ask the question, the information making it true or false might not be in the software at the time. Like … we don't know when the research was published and if there's a new counterargument.*"[42]

It is common for children to overestimate the intelligence of conversational AI, often believing that these systems are smarter than themselves. They assume GenAI should excel at fact-based or computational tasks since it is a computing technology. Learners expressed surprise when a generative AI produced an incorrect answer for a multiplication problem. Even parents hold misconceptions that GenAI platforms perform fact-checking and truth verification before generating responses: "*I believe generative AI can search the internet at an exceedingly high speed, processing numerous queries and capturing the most dominant ones. I assume it uses some form of fact-based checking to respond*"[43]. This suggests that children may be vulnerable to overtrusting conversational AI, particularly since GenAI chatbots can generate convincing information. The inability of GenAI chatbots to consistently produce verified and truthful information makes users prone to misinformation.

Educational activities can reduce initial fears about technologies or understand that GenAI's limitations, particularly its tendency to hallucinate, or make mistakes, highlighting the role of educational activities. Research has shown that exposing children to the errors and limitations of GenAI can mitigate their overtrust. For instance, children's perceptions of machine learning shift after they

---

[39] Flanagan, T., Wong, G., & Kushnir, T. (2023). The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Developmental psychology*, *59*(6), 1017.

[40] The study was conducted with 14 middle school students (12 boys and 2 girls) See more: Kim, K., Kwon, K., Ottenbreit-Leftwich, A., Bae, H., & Glazewski, K. (2023). Exploring middle school students' common naive conceptions of Artificial Intelligence concepts, and the evolution of these ideas. *Education and Information Technologies*, *28*(8), 9827-9854.

[41] The study conducted content analysis on Reddit and interviewed 20 participants (7 teenagers (aged 13-17) and 13 parents) in a local public high school in the US. See more: Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications.

[42] The study explores middle school girls' (N = 26) attitudes and reasoning about how GenAI works in the US, focusing on ChatGPT. See more: Solyst, J., Yang, E., Xie, S., Hammer, J., Ogan, A., & Eslami, M. (2024). Children's Overtrust and Shifting Perspectives of Generative AI.

[43] The study conducted content analysis on Reddit and interviewed 20 participants (7 teenagers (aged 13-17) and 13 parents) in a local public high school in the US. See more: Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications.

participate in training and coding smart programs[44] or learn about the spread of misinformation. These findings highlight the significant role of educational activities in shaping children's understanding of GenAI, promoting a more informed and critical perspective.

1.3. Children's current usage of GenAI-chatbots (UK context)

In general, there has been a rapid increase in awareness and use of GenAI tools and platforms among children in the UK. Awareness of generative AI tools among young people aged 13 to 18 in the UK rose significantly from 20% in early 2023 to 92.2% by early 2024. Usage among those aware of generative AI tools doubled, with 37.1% using these tools in 2023, increasing to 77.1% in 2024[45].

Younger children (under 13 years old) began adopting technology despite restrictions on some generative AI platforms designed for users over 13 in the UK. By 2023, a significant minority of children aged 7-12 (40%) had already started using the technology[46]. Additionally, 79% of online teenagers aged 13-17 were utilizing generative AI tools and services. Among these teenagers, older teens aged 16-18 initially used generative AI more frequently than younger teens aged 13-16. However, by 2024, this trend reversed, with slightly higher usage among teens aged 13-16 compared to those aged 16-18 (77.9% vs. 72.2%)[47].

Even though there is an equal rate of use of GenAI tools by gender, there might be differences in preferences of specific GenAI tools in the UK. In 2023, more boys (40.3%) than girls (23.6%) used generative AI, but by 2024, the usage rates were almost equal (78.3% for boys vs. 76.4% for girls)[48]. However, surveyed teenage girls who are online are the most frequent users of Snapchat My AI, with 75% reporting usage. In contrast, boys are more likely to use ChatGPT more than girls (34% vs. 14%). There is no reported data for non-binary or other gender identities. This trend is similar to that observed in the US, where boys are more than twice as likely to report using ChatGPT (48%) than girls (24%)[49].

There appears to be little difference in generative AI usage between socio-economic groups in the UK. The gap in generative AI usage between those receiving free school meals and those not receiving

---

[44] Druga, S., & Ko, A. J. (2021). How do children's perceptions of machine intelligence change when training and coding smart programs?. In *Proceedings of the 20th annual ACM interaction design and children conference* (pp. 49-61).

[45] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.

[46] Ofcom. (2023). Online nation: 2023 report.

[47] Ofcom. (2023). Online nation: 2023 report.

[48] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.

[49] N=1181 parents with children age 5–18 enrolled in grades K–12 and N=300 students age 12–18 enrolled in grades K–12. See more: Common Sense. (2023). Parents and students are optimistic about AI, but parents have a lotto learn to catch up to their kids – and want rules and ratings to help them.

them narrowed, with both groups having similar usage rates (77.7% vs. 77.3%) by 2024[50]. The initial results might indicate negligible differences among socio-economic groups regarding the use of GenAI; however, more evidence is needed to make a conclusive statement. In contrast, surveys in the US show that teens (13–17 years old) from lower-income households exhibit higher usage of social media and online platforms[51].

Regarding ethnicities, there is limited evidence on the patterns of generative AI usage among children of different ethnic groups in the UK.

Children use a variety of GenAI tools (e.g., text-to-image, text-to-text applications) for diverse purposes, including education, entertainment, experimentation, socialisation, curiosity, and inspiration. Snapchat My AI is the most widely used generative AI tool among online children. By June 2023, half of UK online children aged 7-18 reported using Snapchat My AI. Other popular tools include ChatGPT, Midjourney, or DALL-E. Among internet users aged 16 and above, ChatGPT is the most widely used generative AI service, with 23% reporting usage[52].

Children generally agree that GenAI tools support their learning and creativity; however, there may be a lack of critical engagement with these tools. For schoolwork, students reported using GenAI tools such as ChatGPT, Bard, Midjourney, and Canva[53]. Between 14% and 67% of secondary students have used GenAI for schoolwork and studies[54]. Among children aged 13–18, 44.4% use GenAI for chatting, 18.5% for writing stories, 12.8% for composing poems or lyrics, and 9.0% for writing non-fiction. These tools encourage even those who do not typically enjoy writing to experiment[55]. Younger children aged 8-13 tend to use GenAI more for chatting (58.9%) than for homework (40.7%). They also use it to generate ideas (63.2%), learn new things (56.7%), and read (28.7%) more frequently than older children[56]. However, approximately 20% of surveyed children aged 8–18 admitted to simply copying what GenAI provided or not verifying the information generated by GenAI[57].

Children also use generative AI tools for socialisation and friendship, such as chat assistants in friends' Discord channels. Many have engaged with character-based chatbots to interact with human-like agents on

[50] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.
[51] See reference 26
[52] See reference 23
[53] Department for Education. (2023). Generative AI in education: Educator and expert reviews.
[54] Department for Education. (2023). Generative AI in education Call for Evidence: summary of responses
[55] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.
[56] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.
[57] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.

platforms such as Character.ai. In addition to interacting with these chatbots, some children also create and publish their own character-based chatbots on these platforms[58].

II. OPPORTUNITIES OF GENAI CHATBOTS FOR CHILDREN

2.1. Educational support

Chiu et al. (2023)[59] conducted a systematic review of AI tools in education and found their applications across four key domains: learning, teaching, assessment, and administration. In terms of learning, a review by Zhang et al. (2023)[60] identified how GenAI chatbots facilitate diverse learning activities, including exercises, instructional support, role-playing, collaborative product design, independent writing, storytelling and book reading, digital gameplay, and open-ended debates. Existing studies have shown the potential benefits of GenAI chatbots for students, including increased motivation and engagement, improved academic performance, development of social skills, enhanced quality of instruction and peer communication skills, and support for underserved and vulnerable populations[6162]. Lai et al. (2023)[63] found that intrinsic motivation is the strongest driver for students using chatbots, aligning with prior research that identifies "perceived usefulness" as a key predictor of technology adoption. Early evidence from the Department for Education[64] indicates that teachers across primary, secondary, and tertiary education in the UK acknowledge the benefits and have increasingly used GenAI tools in teaching and learning. However, the presence of a **human expert teacher** remains crucial for providing high-quality education, and the emergence of this technology does not diminish their role in the classroom but supports it.

Personalised learning and assessment: While GenAI shows promise for personalised learning, its application is still in an experimental stage, with the lack of appropriate learning resources being a

[58] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns,Mitigation Strategies, and Design Implications.

[59] Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118.

[60] Zhang, R., Zou, D., & Cheng, G. (2023). A review of chatbot-assisted learning: pedagogical approaches, implementations, factors leading to effectiveness, theories, and future directions. *Interactive Learning Environments*, 1-29.

[61] Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice*, *9*(2), 135-147.

[62] Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118.

[63] Lai, C. Y., Cheung, K. Y., & Chan, C. S. (2023). Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model. *Computers and Education: Artificial Intelligence*, *5*, 100178.

[64] Department for Education. (2023). Generative AI in education Call for Evidence: summary of responses.

significant challenge[65]. Early empirical evidence shows that GenAI chatbots can customise learning materials to fit each student's needs, prior knowledge, learning style, and pace. For elementary students, GenAI tools can aid them in developing reading and writing skills by suggesting syntactic and grammatical corrections or enhancing writing style and critical thinking. These models can also generate questions and prompts that encourage students to analyse and interpret information critically[66]. For instance, Han et al. (2024)[67] explored the use of GenAI sytems, such as ChatGPT and Stable Diffusion[68], in elementary writing education. The study found that GenAI systems can generate adaptive materials based on students' writing abilities, enhance ideation, and provide timely interactions. In this role, GenAI acts more as a coach or peer than assistant, fostering student agency through role allocation. However, concerns remain over authorship and ownership of AI-generated content. Another example is the use of GenAI in creating personalised lessons, illustrations, and exercises tailored to students' knowledge and interests in history subjects in Jauhiainen & Guerra (2023)[69]. For middle and high school students, GenAI tools can support learning in subjects, such as mathematics, physics, and literature, by generating practice problems, quizzes, and step-by-step solutions. Chen et al. (2020)[70] present a list of 30 AI tools used in educational settings, with 70% focused on language learning, 20% on mathematics, and the remainder for other purposes. Some examples of AI tools can be named, such as Amira, Carnegie Learning, Cognii, CueThink and so on. Most of these tools target primary and secondary students, as language and mathematics are key areas of their study. However, the review also highlights a lack of AI tools specifically designed for subjects such as chemistry, literacy, and programming.

Language learning: GenAI chatbots can potentially assist in language acquisition, offering real-time conversation practice, vocabulary building, and grammar correction. A scoping review by Law (2024)[71] and Huang et al. (2022)[72] on the use of GAI tools, including chatbots and large language models (LLMs) in language learning, highlights this as an emerging area, particularly since the launch of ChatGPT in November 2022. Most GenAI applications focus on teaching English, with some extending to other languages such as Turkish, Chinese, Indonesian, and Irish, either as a foreign language (EFL) or second language (ESL). The use of chatbots was reported to decrease shyness and promote confidence among

[65] Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118.

[66] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.

[67] Han, A., Zhou, X., Cai, Z., Han, S., Ko, R., Corrigan, S., & Peppler, K. A. (2024). Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-17)*.

[68] Many contemporary chatbots are powered by ChatGPT and text-to-image generator like Stable Diffusion

[69] Experiments involve 110 pupils, aged 8-14 years old, studying in the 4th–6th grades across four classes in two schools in Uruguay. See more: Jauhiainen, J. S., & Guerra, A. G. (2023). Generative AI and ChatGPT in school children's education: Evidence from a school lesson. *Sustainability*, 15(18), 14025.

[70] Chen, X., Xie, H., & Hwang, G. (2020). A multi-perspective study on Artificial Intelligence in Education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, *1*, 100005.

[71] Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 100174.

[72] Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237-257.

language learners in different skills, including speaking, writing, reading and listening[73]. The use of GenAI spans different educational levels and international tests, including preschool, primary, secondary, higher education, CEFR[74], and TOEFL[75]. For example, Jeon (2023)[76] explored the use of chatbots for Korean primary students learning English and found that chatbots promoted vocabulary acquisition and provided diagnostic information about learner abilities. Other studies show that GenAI tools can suggest alternative word choices, rephrase sentences, and provide real-time feedback, improving learning outcomes and fostering learner autonomy.

Despite concerns about overreliance on GenAI and its potential to hinder critical thinking, GenAI chatbots have the potential to **encourage curiosity and critical thinking** in children by answering questions, posing challenges and promoting problem-solving. These tools guide children to explore topics in depth, ask probing questions, and critically evaluate the information they encounter. For example, Abdelghani et al. (2024)[77] found that GPT-3-generated educational content, particularly open-ended prompts, enhanced children's curiosity-driven question-asking skills more effectively than traditional hand-generated content. However, the study also emphasised that LLMs should complement, not replace, human instruction; should be used under adult supervision due to their 'black-box' nature and unpredictability, and should operate offline to safeguard children's privacy.

While GenAI chatbots show promise in enhancing educational outcomes, **more empirical evidence** is needed to assess their effectiveness and impact. **Ethical concerns** such as academic integrity, bias reinforcement, educational inequality, exposure to inappropriate content, overreliance on AI, and potential negative effects on students' critical thinking and research skills are commonly highlighted in the literature. **Privacy and security risks** are also significant, as the lack of AI literacy among teachers and educators may result in the use of GenAI programs without a clear understanding of data handling practices, increasing the risks of unauthorised access or data breaches involving large-scale data storage. For example, in 2017, a data breach on the Edmodo platform exposed the personal information of over 77 million users[78], and in 2015, Google faced scrutiny when the Electronic Frontier Foundation[79] alleged that Google Apps for Education collected students' data, potentially violating privacy agreements. These cases emphasise the

[73] Experiments involved 68 fifth-grade students in an elementary school in northern Taiwan, aged between 11 and 12, who participated in the reading class to read English books in school. See more: Liu, C. C., Liao, M. G., Chang, C. H., & Lin, H. M. (2022). An analysis of children'interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, *189*, 104576.

[74] The **CEFR** stands for the **Common European Framework of Reference for Languages**. It is an international standard for describing language proficiency. The CEFR framework is widely used in language education to measure and describe learners' language skills across different languages

[75] The **TOEFL** (Test of English as a Foreign Language) is a standardised test used to measure the English language proficiency of non-native speakers. It is widely recognised by universities, colleges, and institutions in English-speaking countries.

[76] Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338-1364.

[77] "Kids Ask" is a prototype for a web-based educational platform that involves an interaction between a child and a conversation agent during a reading-comprehension task. See more: Abdelghani, R., Wang, Y. H., Yuan, X., Wang, T., Lucas, P., Sauzéon, H., & Oudeyer, P. Y. (2024). GPT-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, *34*(2), 483-518.

[78] Twingate. (n.d.). *Edmodo Data Breach: How it Happened & Tips for Protection* (accessed on 25th Oct 2024)

[79] Electronic Frontier Foundation. (2015, December 1). *Google deceptively tracks students' internet browsing, EFF says in complaint to Federal Trade Commission* (accessed on 28th Oct 2024)

importance of strong data protection practices when using educational technologies that involve large-scale data storage. **Intellectual property rights** are another concern, with recommendations that educational institutions could ensure students' work is not used to train GenAI models without proper consent or copyright exemptions. **Longitudinal studies** are necessary to better understand the long-term effects of GenAI tools on educational practices and student learning outcomes[80][81][82][83].

2.2. Creative exploration

Children's creativity inherently differs from that of adults due to their developmental needs and reliance on external sources for cultural and social context[84]. The Human-AI Co-creation model from human-computer interaction (HCI) highlights AI's role as a ***collaborator,*** enhancing users' creative strengths and allowing them to focus on the most creative aspects of their work[85]. Wieland et al. (2022) found that chatbots can be effective partners in human-machine ***brainstorming***, serving as non-judgmental collaborators. Their study found that participants generated more and more diverse ideas when brainstorming with a chatbot compared to a human partner, suggesting that using chatbots can reduce the fear of negative evaluation and enhance idea generation during creative sessions[86]. Furthermore, AI can take on various roles, such as ***Editors, Transformers, Blenders, and Generators***, to facilitate different stages of the creative process[87]. Although research on the relationship between GenAI tools and children's creative processes is still in its early stages, empirical evidence suggests that GenAI tools hold potential to support children's creative expression in storytelling, art, music, writing, and gaming.

Specifically, Newman et al. (2024)[88] show that **GenAI tools can act as constructionist tools for creative self-efficacy**, helping to build children's confidence in their creative abilities by engaging them in

---

[80] McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2024). Generative AI chatbots in higher education: a review of an emerging research area. *Higher Education*, 1-17.

[81] Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 100174.

[82] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

[83] Department for Education. (2023). Generative AI in education: Call for Evidence: summary of responses.

[84] Kudryavtsev, V. (2011). The Phenomenon of Child Creativity. *International Journal of Early Years Education* 19, 1 (2011), 45–53.

[85] Newman, M., Sun, K., Dalla Gasperina, I. B., Shin, G. Y., Pedraja, M. K., Kanchi, R., ... & Yip, J. (2024, May). " I want it to talk like Darth Vader": Helping Children Construct Creative Self-Efficacy with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

[86] Wieland, B., de Wit, J., & de Rooij, A. (2022). Electronic brainstorming with a chatbot partner: A good idea due to increased productivity and idea diversity. *Frontiers in Artificial Intelligence*, *5*, 880673.

[87] Hwang, A. H. C. (2022, April). Too late to be creative? AI-empowered tools in creative processes. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1-9).

[88] Participatory Design with KidsTeam UW involved both adult design researchers (including investigators, master's students, and undergraduates) and 12 child participants. The participants engaged in six design sessions over four months, from February to May 2023. See more: Newman, M., Sun, K., Dalla Gasperina, I. B., Shin, G. Y., Pedraja, M. K., Kanchi, R., ... & Yip, J. (2024, May). " I want it to talk like Darth Vader": Helping Children Construct Creative Self-Efficacy with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

meaningful, hands-on creative experiences. Through these "mini-c" moments (i.e., creativity associated with learning processes), children can gain confidence in their ideas, reinforcing their belief in their ability to complete creative tasks. Drawing on Bandura's concept of self-efficacy[89], shaped by factors such as performance accomplishments, vicarious experiences, verbal persuasion, emotional state, and imaginal experiences, GenAI tools enhance this process by providing immediate feedback and enabling iterative experimentation.

Similar research has suggested prototypes that leverage GenAI tools to support children's creativity. ***StoryPrompt***[90], an interactive system that utilises large language models (LLMs) and AI-generated texts and images (e.g., ChatGPT and Stable Diffusion) to enable *elementary children* to co-create stories and comics. The system combines story and structure planning with emergent storytelling elements, helping to stimulate creativity through the use of words, problem-solving, and visual engagement. The study demonstrates that younger children benefit from GenAI tools by providing ideation resources, while older children face more challenging tasks that encourage reflection, exploration of new possibilities, and potential shifts in narrative direction. Similarly, ***AIStory***[91]**,** an AI-based **visual storytelling application**, leverages generative AI tools such as ChatGPT, Stable Diffusion, and Midjourney to help users create stories, enhancing children's creative expression, storytelling abilities, and literacy development. Another example is ***App Planner***[92], an interactive tool for K-12 students designed to assist in **creating mobile applications**. Utilising GenAI, it guides students through problem-solving and brainstorming via a chat-based interface, helping them generate and refine ideas while stimulating creative thinking. It demonstrates that by integrating creativity with Science, Technology, Engineering, and Mathematics (STEM) learning, GenAI can help transform complex concepts into engaging and accessible experiences, making subjects such as science and math more relatable and fun. Additionally, GenAI can serve as a digital storytelling tool for the **inclusion of students with disabilities**, as demonstrated by Vergara et al. (2024)[93] through voice-based systems designed for the automatic generation of stories targeted at Spanish-speaking children.

These studies collectively emphasise the importance of **participatory design** of incorporating children and relevant stakeholders, including parents, teachers, and caretakers, into the design process of GenAI tools. They also highlight the risks and harms that these tools may unintentionally perpetuate existing biases and stereotypes, highlighting the need for use with **adult supervision**. Those studies also noted the **lack of child-friendly designs**, such as the lack of customisable language options in AI systems, which can be too formal and not tailored to children's domain knowledge or personal interests; or the process of navigating

---

[89] Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, *84*(2), 191.

[90] The study involved interviews with three Chinese teachers on storytelling instruction and focus groups with 18 children (Grades 2-4) to explore their experiences with AI and storytelling preferences. See more: Fan, M., Cui, X., Hao, J., Ye, R., Ma, W., Tong, X., & Li, M. (2024). StoryPrompt: Exploring the Design Space of an AI-Empowered Creative Storytelling System for Elementary Children. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-8).

[91] Han, A., & Cai, Z. (2023, June). Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 470-474).

[92] Kim, D. Y., Ravi, P., Williams, R., & Yoo, D. (2024, June). App Planner: Utilizing Generative AI in K-12 Mobile App Development Education. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference (pp. 770-775).*

[93] Vergara, K. M. R., López-Chau, A., & Hernández, R. R. (2024). Storytelling based on generative AI to promote the inclusion of people with disabilities. *Ingenius*, (32), 101-113.

different platforms, such as ChatGPT and DALL-E, for visual storytelling activities can be confusing for children. More research is needed to understand the **long-term impact of the use of GenAI** tools in promoting creativity in educational settings for children. For example, Kim et al. (2024) raised concerns whether the repetitive and generic answers delivered by those GenAI tools might inadvertently hamper the creativity of students. Doshi & Hauser (2024) suggest that GenAI tools can boost individual creativity but might decrease the novelty and diversity of the content collectively[94]. Therefore, more **longitudinal studies** are needed to examine how GenAI tools influence and potentially support or hinder various aspects of children's development.

2.3. Inclusivity for children with special needs

Inclusive education is an educational approach that values all students equally by removing barriers to learning, enhancing participation, and reducing exclusion within school communities[95]. This approach supports children with special needs or impairments, including those who are *visually impaired*, have *hearing impairments*, *cognitive disabilities*, *motor impairments*, or come from *socio-economically disadvantaged backgrounds*[96]. Some disabilities are quite common, such as dyslexia, the most common specific learning disorder (SLD), which affects 5% to 15% of the global population and impacts reading, writing, and sometimes spoken language comprehension[97]. Individuals with such disabilities often face challenges integrating into society and may experience stigma, isolation, and bullying, which can lead to severe outcomes, including suicide attempts. GenAI chatbots can play a role in supporting these children by offering personalised learning experiences, assisting in the development of communication skills, and promoting confidence and autonomy. AI-powered voice recognition can also assist students with impairments to enhance accessibility.

GenAI chatbots can offer **personalised learning** to meet the individual needs of children with disabilities**,** such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), and dyslexia. These chatbots can adapt educational content to match the child's learning pace, style, and specific challenges. Barua et al. (2022)[98] reviewed the range and effectiveness of AI-based assistive tools designed to support students with neurodevelopmental disorders, who often experience social and communication challenges that can affect their learning. The study found that specific AI tools have been successful in aiding students with reading comprehension, phonics, writing, spelling, handwriting, and mathematics

[94] Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), eadn5290.

[95] Šumak, B., López-de-Ipiña, D., Dziabenko, O., Correia, S. D., de Carvalho, L. M. S., Lopes, S., ... & Pušnik, M. (2024). AI-Based Education Tools for Enabling Inclusive Education: Challenges and Benefits. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 472-477). IEEE.

[96] Šumak, B., López-de-Ipiña, D., Dziabenko, O., Correia, S. D., de Carvalho, L. M. S., Lopes, S., ... & Pušnik, M. (2024). AI-Based Education Tools for Enabling Inclusive Education: Challenges and Benefits. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 472-477). IEEE.

[97] American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders. Fifth Edition. Arlington, VA.

[98] The research reviewed the publications from 2011 to 2021. See more: Barua, P. D., Vicnesh, J., Gururajan, R., Oh, S. L., Palmer, E., Azizan, M. M., ... & Acharya, U. R. (2022). Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. *International Journal of Environmental Research and Public Health*, *19*(3), 1192.

computation. One characteristic that enable AI-based tools to support children with mental disorders (e.g. autism spectrum disorders) is their predictability and constancy. Another example is **AI4LA**[99], a web intelligent chatbot is designed to support students with Dyslexia and Specific Learning Disorders (SLD) by modeling students' understandings and misconceptions for customised educational assistance. The chatbot is also able create concept maps from conversations to visualise the student's knowledge, aiding students in their learning.

Children with **social communication challenges,** such as those on the autism spectrum, can benefit from interacting with GenAI chatbots that simulate social scenarios to improve communication skills. These chatbots can model appropriate social behaviors, such as taking turns in conversation, recognising emotions, and responding to social cues. For example, Tang et al. (2024) developed an AI-based tool, **EmoEden,** to support children with High-Function Autism[100] in **emotional learning**. This tool provides personalised assistance to children aged 8 to 12, helping them identify and express their own emotions while also improving their ability to recognise and respond to the emotions of others[101]. Another example is "**CapacitaBOT**"[102], a chatbot-based mobile application aimed at helping people with intellectual disabilities enhance their communication skills through voice and text-based features. By facilitating automated conversations between the user and the machine, the chatbot prepares users for real-life situations and serves as a resource for acquiring social skill and improving social interaction.

GenAI chatbots can be tailored to support diverse learning needs in the classroom, particularly for students who are disadvantaged, come from varied backgrounds, or have different learning styles. The "Sammy" chatbot, as explored by Gupta et al. (2022)[103], has shown potential in creating an inclusive learning environment. It can answer basic course questions, connect students with campus resources, provide supplementary materials for advanced learning, and offer support for life and wellbeing issues. These benefits stem from the chatbot's accessibility, interactivity, and ability to link students with external resources.

While there are multiple prototypes and studies exploring the potential of chatbots for these children, it remains unclear to what extent these applications have been implemented in practice. Torrado et al., (2023) provides a scoping review on chatbots for children with special education needs and highlights that research in this area is still in its early stages. There is a knowledge gap regarding the effectiveness of commercial chatbots when used for their intended purpose with these individuals. This suggests a need for further research to examine how repurposing commercial chatbots and developing customised chatbots can

[99] D'Urso, S., & Sciarrone, F. (2024, June). AI4LA: An Intelligent Chatbot for Supporting Students with Dyslexia, Based on Generative AI. In *International Conference on Intelligent Tutoring Systems* (pp. 369-377). Cham: Springer Nature Switzerland.

[100] High-Functioning Autism (HFA) individuals have normal intelligence but face challenges in forming social relationships due to difficulties in recognising and expressing emotions.

[101] Tang, Y., Chen, L., Chen, Z., Chen, W., Cai, Y., Du, Y., ... & Sun, L. (2024, May). EmoEden: Applying Generative Artificial Intelligence to Emotional Learning for Children with High-Function Autism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-20).

[102] Melo, A. C., Blanco, L. S., & García, A. M. (2022). Chatbot, as Educational and Inclusive Tool for People with Intellectual Disabilities. *Sustainability*, *14*(3), 1520.

[103] Chatbot called "Sammy" was developed to act as an intelligent and inclusive tutor as part of a user study conducted with 215 undergraduate students in the United States. See more: Gupta, S., & Chen, Y. (2022). Supporting inclusive learning using chatbots? A chatbot-led interview study. *Journal of Information Systems Education*, *33*(1), 98-108.

better support individuals with special needs. The review also emphasises the importance of involving end-users in the co-designing process, as the current approach mostly involves end-users merely as experimental subjects for testing the chatbot[104].

There are also concerns and challenges related to the development and use of AI-driven tools for children with special education needs: (1) Data scarcity: Collecting suitable datasets is challenging due to the complexities of neurodevelopmental disorders (NDDs) and their comorbidities, which limits the development of personalised AI tools. For example, feedback from teachers indicates that tools such as ChatGPT struggle to produce high-quality, detailed content tailored to the diverse needs of disabled students[105]; **(2) Ethical issues**: The use of AI in educational settings raises concerns about privacy, data security, and informed consent, particularly for children with special needs who might have limited capacities to consent. Additionally, there are concerns about the potential negative impact of AI tools on children, for example, hallucinations and bias on high-functioning autistic (HFA) children[106]**; (3) Lack of involvement in the co-design process**: The current approach mostly involves end-users, including children, parents and caretakers as experimental subjects for testing the chatbot[107]. There is a need to have inputs from children and multistakeholders to ensure the tools meet the special needs[108]. While showing potentialities in assisting children with special needs, the tool might also risk exclusion rather than inclusion if not use responsibly.

2.4. Online safety/Child protection

GenAI chatbots can serve as essential tools for detecting, preventing, and intervening in online abuse. Online abuse manifests in various forms, including (1) *cyberbullying*, where children may encounter threatening messages, embarrassing or distressing images or videos, or be harassed on social networks; (2) *sexting*, where children are pressured or coerced into producing sexual images; and (3) *online grooming*, where offenders develop a trusting relationship for sexual exploitation purposes. According to a report by Ofcom, 34% of UK children in the survey have experienced someone being nasty or hurtful to them personally, while 31% have reported harassment through digital communication channels. Notably, 32% of girls receiving unsolicited explicit images (compared to 11% of boys) and 24% being asked to share such

[104] Torrado, J. C., Bakke, C., & Gabarron, E. (2023). Chatbots and Children with Special Educational Needs Interaction. In *International Conference on Human-Computer Interaction* (pp. 443-452). Cham: Springer Nature Switzerland.

[105] Seiradakis, E. V. (2023). Unpacking Experts' Opinions on ChatGPT Potential Assistive Roles and Risks in Early Childhood Special Education. In *International Conference on New Media Pedagogy* (pp. 380-392). Cham: Springer Nature Switzerland.

[106] Tang, Y., Chen, L., Chen, Z., Chen, W., Cai, Y., Du, Y., ... & Sun, L. (2024, May). EmoEden: Applying Generative Artificial Intelligence to Emotional Learning for Children with High-Function Autism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-20).

[107] Torrado, J. C., Bakke, C., & Gabarron, E. (2023). Chatbots and Children with Special Educational Needs Interaction. In *International Conference on Human-Computer Interaction* (pp. 443-452). Cham: Springer Nature Switzerland.

[108] Barua, P. D., Vicnesh, J., Gururajan, R., Oh, S. L., Palmer, E., Azizan, M. M., ... & Acharya, U. R. (2022). Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. *International Journal of Environmental Research and Public Health*, *19*(3), 1192.

images of themselves (compared to 9% of boys)[109], highlighting the vulnerabilities of children faced on online environments.

GenAI chatbots can be used to monitor real-time online interactions, detect and flag instances of abuse, and identify potential offenders. For example, **Protectbot**[110] is a conversational AI model that uses conversation analysis for the real-time detection and mitigation of potential sexual predatory behaviors in gaming environments. Similarly, **BotHook**[111] is an AI-based platform designed to characterise trends among cyber predators and act as a trap for online criminals. This distributed chatbot platform includes a module to attract pedophile interest, an intelligent engine for question-answer analysis, and an automated system for identifying and profiling pedophile behavior patterns. Another example is **Negobot**[112], a conversational agent that poses as a child to engage with potential offenders. Using game theory, Negobot navigates conversations to gather information that may indicate pedophile tendencies for effective identification of potential predators.

GenAI chatbots can act as preventive tools to help deter criminal activities before they escalate by providing early intervention. For example, the **reThink** chatbot[113], developed by the Internet Watch Foundation (IWF) and Stop It Now!, has been launched on websites such as Pornhub to deter potential offenders attempts to search for child sexual abuse material. The chatbot provides users with educational information on the harm caused by such behaviours and directs them to resources for help and support. Another example is the **Sweetie 2.0 chatbot**[114], a virtual 10-year-old Filipino girl designed to identify potential child sex offenders in chatrooms, particularly in live-streamed webcam scenarios where little digital evidence is left for investigation. With an automated chat function, Sweetie 2.0 can track, locate and identify these offenders and send them warning messages of potential legal consequences.

GenAI chatbots can also be served as educational tools to raise awareness among young people about online threats. For example, **Seri** (Stop cybERgroomIng)[115] chatbot framework helps young users recognise and respond to online threats by simulating realistic conversations between a perpetrator chatbot and a potential victim chatbot in a safe and controlled environment. By interacting with the chatbot, children can become more aware of potential dangers, learn to identify predatory behavior, and practice appropriate responses (e.g., when being asked for private or sensitive information). This knowledge is particularly important for children, as studies show that 12-year-olds have a significantly lower comprehension of

[109] Ofcom. (2024). Childrens Media Literacy Report.

[110] Faraz, A., Ahsan, F., Mounsef, J., Karamitsos, I., & Kanavos, A. (2024). Enhancing Child Safety in Online Gaming: The Development and Application of Protectbot, an AI-Powered Chatbot Framework. *Information*, *15*(4), 233.

[111] Zambrano, P., Sanchez, M., Torres, J., & Fuertes, W. (2017, October). BotHook: An option against cyberpedophilia. In *2017 1st Cyber Security in Networking Conference (CSNet)* (pp. 1-3). IEEE.

[112] Laorden, C., Galán-García, P., Santos, I., Sanz, B., Nieves, J., Bringas, P. G., & Gómez Hidalgo, J. M. (2015). Negobot: Detecting paedophile activity with a conversational agent based on game theory. *Logic Journal of IGPL*, *23*(1), 17-30.

[113] Internet Watch Foundation (IWF). (2022). reThink Chatbot (accessed 26th August 2024)

[114] Terre des Hommes. Sweetie.

[115] Wang, P., Guo, Z., Huang, L., & Cho, J. H. (2021). Seri: Generative chatbot framework for cybergrooming prevention.

sexual content in digital media compared to adolescents aged 14 to 16 years, making them less likely to recognise warning signs[116].

GenAI chatbots can offer support to children who are victims of online abuse by guiding them on how to respond and encouraging them to seek help from trusted adults. For example, Piccolo et al. (2021)[117] conducted a study with 110 schoolchildren in the UK, using LEGO figures to investigate the potential of chatbots in helping children facing online threats. These chatbots could serve as an entry point to child protection services, providing both emotional and practical support for those preparing to seek help or report a crime. The chatbots are designed to enhance communication between children and their existing support networks, including parents, schools, friends, and the broader community. By involving children in the design process through participatory methods, the study emphasises the importance of incorporating children's perspectives, hence empowering them throughout the development of these supportive tools. Another example is **SomeBuddy[118]**, an AI-based chatbot that provides legal and psychological advice to children and adolescents who have potentially experienced online harassment, acting as a first-aid kit for the affected child.

Additionally, GenAI can contribute to creating a safer online environment in general. For example, GenAI can be employed for **hate speech detection[119]**, identifying and flagging harmful language. Moreover, GenAI can also be instrumental in the **early detection of cyber threats[120]** such as scams or phishing attempts by analysing patterns in online behavior and communications, allowing for timely interventions. GenAI can also support **content moderation[121]** by automatically identifying and removing inappropriate materials, such as CSAM.

However, using GenAI for children's digital safety also presents potential risks and harms that require careful consideration. These include **(1) privacy and security risks** related to collecting, handling, and storing sensitive data about children; **(2) the risk of providing inaccurate or harmful advice**, particularly in mental health scenarios involving vulnerable children; and **(3) ethical and legal issues**, such as those

[116] Brown, J. (Ed). (2008). Managing the Media Monster: The Influence of Media (From Television to Text Messages) on Teen Sexual Behavior and Attitudes. Washington, DC: National Campaign to Prevent Teen and Unplanned Pregnancy.

[117] The study conducted eight design sessions with 110 schoolchildren aged 11 to 17 in UK primary and secondary schools between October and December 2019. See more: Piccolo, L. S. G., Troullinou, P., & Alani, H. (2021). Chatbots to support children in coping with online threats: Socio-technical requirements. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1504-1517)

[118] UNICEF. (2021). Global Insight AI case study SomeBuddy.

[119] Skarzynska, E., & Paliszkiewicz, J. The Use of Generative Artificial Intelligence (GenAI) Capabilities for Early Detection of Threats in the Digital Environment: The Good Side of GenAI. In *Regulating Hate Speech Created by Generative AI* (pp. 91-104). Auerbach Publications.

[120] Sai, S., Yashvardhan, U., Chamola, V., & Sikdar, B. (2024). Generative AI for cyber security: Analyzing the potential of chatgpt, dall-e and other models for enhancing the security space. *IEEE Access*.

[121] Laranjeira da Silva, C., Macedo, J., Avila, S., & dos Santos, J. (2022, June). Seeing without looking: Analysis pipeline for child sexual abuse datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2189-2205).

raised by the "Sweetie" case, which questions the legality of using chatbots as investigative tools and whether their use might constitute entrapment[122].

2.5. Parenting, teaching and professional support

This section will explore how GenAI chatbots can be used to support parents, teachers and professionals whose work concerns children.

Parenting support: While much research has focused on the use of GenAI tools in educational settings, few studies have explored their application in informal learning environments such as the home, which is critical for children's development. Home learning plays an important role in shaping children's school readiness and academic performance[123]. GenAI chatbots, such as ChatGPT, can assist parents by providing guidance on child development, behavior management, and educational activities tailored to a child's age and needs. Research has shown that parents have used GAI tools such as ChatGPT to **seek information on topics such as autism**[124]. While the AI provided largely accurate, concise, and clear information, it lacked actionable advice and contained inaccurate references and hyperlinks. In other contexts, GenAI tools have been used to help parents develop **effective praising skills** and explore parental characteristics, such as assumptions about parenting[125]. More research is needed to further explore the potential and limitations of GenAI tools in supporting parenting at home.

Teaching support: GenAI chatbots can assist teachers in various tasks such as lesson planning, assessment, professional development and administrative tasks. Evidence from teachers in the UK shows that teachers have already used GenAI chatbots to assist in lesson planning by generating lesson plans, syllabi, and activities[126]. GenAI chatbots can semi-automate the grading process, especially in disciplines such as language, writing, speaking and mathematics, with more accurate and faster assessments for effective grading[127]. GenAI tools can support teachers' professional development by offering real-time feedback on their teaching practices. AI agents can analyse classroom data, such as behaviours and questioning skills, and provide constructive suggestions for improvement[128]. GenAI tools can also support educational

---

[122] van der Hof, S., Georgieva, I., Schermer, B., Koops, B. J., & Tourism, W. C. S. (2019). Sweetie 2.0. *Information Technology and Law Series*, *31*.

[123] Quan, S., Du, Y., & Ding, Y. (2024, May). Young Children and ChatGPT: Parents' Use of ChatGPT in Parenting. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

[124] McFayden, T. C., Bristol, S., Putnam, O., & Harrop, C. (2024). ChatGPT: Artificial Intelligence as a Potential Tool for Parents Seeking Information About Autism. *Cyberpsychology, Behavior, and Social Networking*, *27*(2), 135-148.

[125] 170 parents reside in Argentina and have at least one child between the ages of 2 and 11 years old. See more: Entenberg, G. A., Mizrahi, S., Walker, H., Aghakhani, S., Mostovoy, K., Carre, N., ... & Bunge, E. L. (2023). AI-based chatbot micro-intervention for parents: Meaningful engagement, learning, and efficacy. *Frontiers in Psychiatry*, *14*, 1080770.

[126] Department for Education. (2023). Generative AI in education Call for Evidence: summary of responses.

[127] Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118.

[128] Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(5), 87-101.

decision-making by providing evidence-based information such as the likelihood of students discontinuing their courses, identifying factors affecting academic performance, and assisting with course selection[129].

Professional support: GenAI chatbots can support professionals working with children, including paediatricians, child psychologists, social workers, law enforcement officials, and child protective services (CPS) personnel. Beyond handling operational tasks such as administrative duties and scheduling appointments, GenAI tools can assist in sensitive areas, such as training professionals to interview children in cases of abuse. These interviews are not only essential for gathering crucial information but must also be conducted in a way that avoids causing secondary trauma and helps children feel supported. Techniques such as asking open-ended questions and avoiding leading ones are key to a successful interview, but mastering these skills requires proper training. Early prototypes of GenAI-powered child avatars[130] have demonstrated the potential for enhancing interview skills by allowing professionals to practice and refine their techniques in a controlled, realistic setting.

III. POTENTIAL RISKS/HARMS OF GENAI CHATBOTS FOR CHILDREN

The harms associated with GenAI chatbots for children should be understood on multiple levels, including **individual, community, and societal impacts**. At the **individual level**, the technology can affect children directly, influencing their mental health, emotional well-being, cognitive development, or behaviour. At the **community level**, the impact extends beyond individuals to the groups they are part of, such as families, peer networks, schools, or local communities. Here, chatbots could disrupt traditional relationships, alter communication dynamics, or influence cultural and social norms within these smaller networks. Lastly, at the **societal level**, the influence becomes more systemic, shaping collective values, norms, or policies. This could include concerns such as perpetuating biases at scale, reinforcing societal inequalities, or influencing public trust in technology. Acknowledging these harms across these interconnected levels is essential to comprehensively assess the risks and benefits of GenAI chatbots and address their implications effectively.

Given the wide range of effects these technologies can have, it is important to consider how GenAI chatbots influence children's development across various domains, including **physical, societal, emotional, psychological, and cognitive aspects**. These impacts can manifest in both the **short term and long term**, affecting a child's overall growth and well-being.

[129] Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education*, *17*, 1-13.

[130] Salehi, P., Hassan, S. Z., Lammerse, M., Sabet, S. S., Riiser, I., Røed, R. K., Johnson, M. S., Thambawita, V., Hicks, S. A., Powell, M., Lamb, M. E., Baugerud, G. A., Halvorsen, P., & Riegler, M. A. (2022). Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children. *Big Data and Cognitive Computing*, *6*(2), 62.

Table 1: Summary of risks/harms of GenAI chatbots for children

Sources of harm	Types of outputs	Features	Online harms	Examples
The intrinsic nature of technology	Age-inappropriate	Content not suitable for children's age	Emotional distress, inappropriate knowledge, or behavior	Sexually explicit material, violent imagery
	Toxic/harmful contents	Content promoting hate, violence, or cruelty	Encouragement of harmful behavior, negative psychological impact	Hate speech, cyberbullying, animal cruelty, substance abuse, and self-harm
	Inaccurate information (misinformation)	False or misleading information	Development of misconceptions, poor decision-making	Incorrect historical facts, dangerous health advice
	Biased contents	Content reflecting bias or prejudice	Reinforcement of stereotypes, development of biased views	Gender or racial stereotypes, discriminatory language
	Emotional and psychological manipulation	Manipulative or emotionally charged content	Emotional distress, unhealthy attachment to AI, mood swings	AI simulating emotional responses to manipulate feelings
Structural	Privacy and security	Collection and sharing of personal data	Risk of data breaches, identity theft, misuse of information	Storage of children's personal information without consent
Misuse of technology	Overreliance	Excessive dependence on AI-generated responses	Reduced critical thinking, loss of problem-solving skills, affecting cognitive development	Constantly asking AI for answers instead of thinking independently
	Addiction	Compulsive use of AI technologies	Neglect of real-world interactions, affect social skills, negative impact on mental health	Spending excessive time interacting with AI chatbots

Table 2 outlines the complex risks associated with the deployment of generative AI chatbots for children across four key domains: technology, children's developmental stages, ecosystems, and regulatory frameworks.

Table 2: Risks factors

Technology	Children	Ecosystems	Regulations
Hallucination Bias Data collection Personalisation Human-like	Immature physical, psychological, emotional, cognitive, development	Lack of understanding of technology and oversight of children's use	Operationalisation issues Lack of child participation Technical challenges

Technology: Several risks arise from the inherent nature and misuse of generative AI. First, **hallucination** occurs when AI models generate incorrect or nonsensical information. This poses a significant risk for children, who may accept such information as fact, potentially leading to misinformation or confusion. **Bias** is another critical issue, as AI systems trained on vast datasets can inherit and amplify societal prejudices, potentially exposing children to harmful stereotypes or discriminatory content. Additionally, the reliance on **data collection** for personalized interactions raises privacy concerns, especially with the sensitive data of young users. **Personalisation** itself, while beneficial for engagement, can lead to over-dependence on AI, limit exposure to diverse perspectives, and make it difficult to regulate content appropriateness for children. Finally, the **human-like interaction** offered by some AI systems can blur the line between technology and real human relationships, potentially causing children to form attachments or misunderstand the non-human nature of AI. Additionally, when children perceive chatbots as more human-like, they are more likely to share sensitive information, potentially increasing privacy and security risks.

Children: The child-specific risks are rooted in the unique developmental vulnerabilities of children. Children's unique perspectives and limited life experiences shape how they engage with AI and interpret its outputs. Due to their ongoing emotional, cognitive, and social development, they are more susceptible to risks such as manipulation, bias, and data privacy breaches[131]. They may lack the maturity to critically evaluate AI responses or understand the limitations and artificial nature of the technology. Additionally, children are still forming their identities and are more influenced by social comparisons and peer validation, which GenAI can exacerbate. AI systems, by amplifying negative biases, can have long-lasting effects on children's self-esteem, mental health, and identity development[132]. Moreover, with limited impulse control and emotional regulation, children may lack the autonomy and decision-making capacity necessary to navigate these technologies safely. This immature development increases the risk of negative impacts, as children may internalise inappropriate content or become overly reliant on AI without fully understanding its limitations.

Ecosystems: Children's use of GenAI tools is often beyond parents' and teacher's understanding and knowledge. Though there is limited evidence worldwide, initial studies indicate that children have used

[131] Reich, S. M., Starks, A., & Su, Z. (2024). One (Adult) Size Does Not Fit All: The Importance of Development in Digital Design and Utilization. *Youth Wellbeing in a Technology Rich World*.
[132] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.

more GenAI than adults[133]. Parents have a limited understanding of children's use of generative AI tools, reporting that their children never used such tools or only used ChatGPT[134]. In the UK, 37% of parents of secondary school students were unsure whether their child used AI for schoolwork[135]. Students may be using and even more familiar with GenAI tools than their teachers[136]. Similarly, in the US, while 50% of students aged 12-18 say they have used ChatGPT for school, only 26% of parents in that age range are aware of it. Additionally, 38% of students admit to using ChatGPT for a school assignment without their teacher's permission, and 56% say they know a friend or classmate who has done the same[137].

Regulations: Operationalisation issues make it difficult to translate child safety regulations into practical, enforceable standards, especially given the rapid pace of AI advancements. There is also a lack of child participation in the development of regulatory frameworks, resulting in policies that may not fully address the unique needs and concerns of young users. Additionally, technical challenges arise in keeping regulatory standards up-to-date with technological innovations, as well as in addressing persistent issues like data privacy and bias mitigation within AI systems.

3.1. Exposure to age-inappropriate, toxic and harmful contents

Under the Online Safety Act, platforms are responsible for safeguarding children from explicit content, including pornography and material that promotes or instructs on self-harm, eating disorders, or suicide, classified as *Primary Priority Content*. Additionally, platforms must ensure that children are exposed only to age-appropriate *Priority Content*, which includes harmful material such as bullying, abusive or hateful language, depictions of serious violence or injury, dangerous stunts or challenges, and content that encourages the ingestion, inhalation, or exposure to harmful substances[138].

Generative AI chatbots, trained on large, unchecked datasets scraped from the open internet, are prone to generating harmful, offensive, or age-inappropriate content[139]. Contemporary GenAI chatbots are either built from scratch with large corpora or fine-tuned from pre-trained models such as OpenAI's GPT and Meta's LLaMA. Either way, these datasets often include toxic material such as hate speech, cyberbullying, and misleading information, which can lead to chatbots producing unsafe responses. The variety of toxic content can be categorised as offensive, targeted or non-targeted, individual (cyberbullying), group (hate speech) or others[140]. There have been numerous instances in which chatbots

[133] UNICEF. (2023). Generative AI: risks and opportunities for children. UNICEF Innocenti–Global Office of Research and Foresight.
[134] See reference 37
[135] Pupils in years 7 to 13 (n=3,238), secondary parents (n=1,738). See more: Department for Education. (2023). Parent, Pupil and Learner Panel (April/May 2023).
[136] Department for Education. (2023). Generative AI in Education: Educator and expert reviews.
[137] N=1181 parents with children age 5–18 enrolled in grades K–12 and N=300 students age 12–18 enrolled in grades K–12. See more: Common Sense. (2023). Parents and students are optimistic about AI, but parents have a lotto learn to catch up to their kids – and want rules and ratings to help them.
[138] Department for Science, Innovation and Technology. (2024). Guidance **Online Safety Act: explainer.**
[139] Chua, J., Li, Y., Yang, S., Wang, C., & Yao, L. (2024). AI Safety in Generative AI Large Language Models: A Survey.
[140] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1.*

were suspended due to their harmful, racist or toxic content generation (e.g., Microsoft's TwitterBot Tay, Luda chatbot).

The risks are heightened by the fact that some platforms lack robust age verification or filtering systems, making it easier for children to access content that is not age-appropriate, harmful or offensive. For example, participants in Yu et al. (2024)[141] noted that children can request explicit content, such as sexual and violent material, on GenAI chatbots (e.g., Chai), *"Character AI is the one most people are famili*ar with, but Chai has fewer restrictions. Chai is basically a hotspot for NSFW (not safe for work) bots.". For instance, in a safety test, Snapchat's My AI chatbot advised a pretended 13-year-old girl on how to use candles and music when losing their virginity to a 31-year-old partner without recognising a potential child abuse situation[142]. In another case, when a user mentioned that their parents wanted to delete the Snapchat app, My AI suggested having an honest conversation with them but offered deceptive advice on hiding the app on another device, thus bypassing parental supervision[143].

Exposure to age-inappropriate, harmful and offensive content on GenAI chatbots can have long-lasting impacts on children, affecting their psychological, emotional, and social development. **However, there is a lack of comprehensive longitudinal studies on the impact of GenAI chatbots on children, leaving critical research gaps that require further exploration**. Such exposure can lead to emotional distress, anxiety, and fear, while repeated encounters with violence or toxic material may desensitise children, making them more likely to normalise harmful behaviours (e.g., increased aggression, violent radicalization, eating disorders, suicides) or develop risky behaviours (e.g., drug use, self-harm, unsafe sexual practices, physical stunts, online risk-taking)[144][145][146]. It can also distort their perceptions of reality, affecting their understanding of relationships, body image, and societal norms. The cumulative effect during childhood can lead to long-term mental health challenges and social and cognitive issues in adulthood. Given their still-developing brains, children may struggle to comprehend the consequences of their actions and are more likely to be attracted to harmful content[147]. As GenAI systems are increasingly designed to increase user trust, human likeness, and personalisation, they even make children more susceptible to harmful suggestions and less likely to take protective actions, such as reporting or disengaging from harmful content[148].

---

[141] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications.

[142] Fowler, G. (2023). Snapchat Tried to Make a Safe AI. It Chats with Me about Booze and Sex. Washington Post.

[143] Fowler, G. (2023). Snapchat Tried to Make a Safe AI. It Chats with Me about Booze and Sex. Washington Post.

[144] Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., ... & Sieckelinck, S. (2018). Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International journal of developmental science*, *12*(1-2), 71-88.

[145] Chiang, J. T., Chang, F. C., & Lee, K. W. (2021). Transitions in aggression among children: Effects of gender and exposure to online violence. *Aggressive behavior*, *47*(3), 310-319.

[146] Brennan, C. (2024). AI being used to harm children online, committee to hear today. Irish Examiner (accessed 27th September 2024).

[147] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.

[148] Ofcom. (2023). Online nation: 2023 report.

3.2. Misinformation/disinformation harms

Misinformation refers to false or inaccurate information that may be spread unintentionally without the deliberate intent to cause harm[149]. On the other hand, disinformation is defined as "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit"[150]. Thus, while disinformation involves a deliberate intent behind the dissemination of false information, misinformation can be shared without the intent to deceive. Nonetheless, both misinformation and disinformation can lead to significant harm, including misleading, deceiving, or manipulating individuals, inciting violence, exacerbating social polarisation, targeting minority groups, undermining democratic processes, and eroding trust within communities[151][152]. With their developing cognitive abilities and limited experience, children are especially vulnerable to the deceptive nature of false information, which can have serious consequences for their mental, emotional, and social development.

GenAI has amplified the reach and impact of misinformation[153][154]. AI chatbots and large language models (LLMs) can act as potential amplifiers of misinformation by generating false or factually incorrect responses (also known as hallucinations) by assigning high probabilities to false or misleading statements[155]. GenAI chatbots can also be used to create persuasive deepfake materials on text, images, audio, and video, blurring the line between authentic and fabricated content[156]. Evaluation by Lin et al. (2022)[157] on LLMs found that the larger models tend to be the least truthful, with even the best-performing model being accurate in only 58% of cases, compared to 94% accuracy in human performance. Another example is Google's Bard, which was found to generate misinformation without disclaimers for 78 out of 100 false and potentially harmful narratives on topics such as climate change, vaccines, Lhomophobia and transphobia, and sexism[158].

---

[149] George, J. F. (2024). Discovering why people believe disinformation about healthcare. *PLOS ONE*, *19*(3), e0300497.

[150] HLEG. A multi-dimensional approach to disinformation: report of the independent high level group (HLEG) on fake news and online disinformation. *European Commission* 2018. Publications Office of the European Union.

[151] Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298-320.

[152] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models.

[153] Mitra, A., Mohanty, S. P., & Kougianos, E. (2024). The World of Generative AI: Deepfakes and Large Language Models.

[154] Williams, A. (2024). Online misinformation: how generative AI and LLMs are changing the game. The Alan Turing Institute.

[155] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models.

[156] The study utilised three models of GPT2 for their experiments. See more: Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, *9*(1), 104-117.

[157] Evaluation was conducted on We tested GPT-3, GPT-Neo/J, GPT-2 and a T5-based model. See more: Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

[158] Center for Countering Digital Hate. (2023). Bard: Google's new AI chat generates misinformation when prompted on 78 out of 100 false and potentially harmful narratives without disclaimers.

As more young people turn to social media as their primary source of information, they are increasingly exposed to misinformation presented on those sites. Despite that, they struggle to detect deep fakes and fake news and lack reliable methods to verify the accuracy of the information. Misinformation is particularly prevalent on social media, with 78% of UK adults reporting encounters with false information and 50% having received it directly[159]. Another survey found that 68% of UK users had seen deliberate misinformation online, and the true extent is likely even higher, as many individuals fail to recognise misinformation when they see it[160]. Despite the high prevalence of untrustworthy information on social media, young people continue to rely on these platforms—many of which integrate Generative AI (GenAI) tools—as their dominant news sources. In particular, for children aged 12-15, TikTok is the most-used news source at 28%, while for those aged 16–24, Instagram leads with 44% usage[161].

Although children may have varied perceptions about the accuracy of the information they encounter and might be sceptical of it, their scepticism is often shallow and can be overridden by the deceptive nature of misinformation[162]. Several experiments have shown that children struggle to reliably detect deepfakes, even after educational training activities (e.g., Ali et al., 2021)[163]). They also have limited capacities to detect fake news (e.g., identify the spoof website "Save the Pacific Northwest Tree Octopus"[164][165]). This could be attributed to children's limited exposure to manipulated media, a lack of contextual knowledge, and the increasingly sophisticated methods used to make such content appear real[166]. Despite these challenges, children often overestimate their ability to discern authentic content from manipulated media and generally lack effective methods to verify its accuracy[167], making them more vulnerable to the harmful effects of misinformation[168].

[159] Based on surveys conducted in March 2023 with 1,993 nationally representative participants aged 18-88 from the UK across genders, and ethnicities. Full report: The Alan Turing Institute. (2024). How do people protect themselves against online misinformation.

[160] Based on surveys conducted in June 2023 with 4,150 nationally representative consumers aged 16-75 in the UK weighted for demographics such as age, gender, region, and working status. Full report: Deloitte. (2023). Digital Consumer Trends 2023.

[161] A total of 501 interviews with individuals aged 12-15 and 4,556 interviews with adults were conducted in 2022 and 2023. Full report: Ofcom. (2023). News consumption in the UK.

[162] Shtulman, A. (2023). Children's susceptibility to online misinformation. *Current Opinion in Psychology*, 101753.

[163] Experiments were conducted with 38 middle-school students from five states across the United States. See more: Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence*, *2*, 100040.

[164] In this experiment in the Netherlands, only 2 out of 27 school children (7 per cent) were able to identify the website as a hoax. See more: Loos, E., Ivan, L., & Leu, D. (2018). "Save the Pacific Northwest tree octopus": a hoax revisited. Or: How vulnerable are school children to fake news?. Information and Learning Science, 119(9/10), 514-528.

[165] The experiments were conducted with primary and secondary school students in the US. See more: Pilgrim, J., Vasinda, S., Bledsoe, C., & Martinez, E. (2019). Critical thinking is critical: Octopuses, online sources, and reliability reasoning. *The Reading Teacher*, *73*(1), 85-93.

[166] Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. Computers and Education: Artificial Intelligence, 2, 100040.

[167] Ofcom. (2024). Children and Parents: Media Use and Attitudes Report.

[168] Ofcom. (2024). Children and Parents: Media Use and Attitudes Report.

Children can overtrust or overestimate the capabilities of GenAI chatbots, making them vulnerable to being misled or manipulated. As highlighted in Section II, many children overestimate the intelligence of GenAI chatbots, often believing these systems are smarter than they are. In a survey conducted by the UK Department of Education, approximately 20% of children aged 8–18 admitted to simply copying information provided by GenAI or not verifying the accuracy of the generated content, demonstrating a lack of critical engagement[169]. This tendency can lead children to accept chatbot responses as truth, resulting in misunderstandings about critical topics such as health, safety, and social issues, and can actually cause harm in many cases. For example, misinformation on medical dosages may lead a user to cause harm to themselves[170]. Misinformation on sensitive topics such as immigration, gender, politics and equality can exacerbate social polarisation among children by fostering divisions based on false narratives or manipulated facts[171]. As children begin forming opinions about social and political issues in their formative years, particularly during middle and high school, consuming misinformation can solidify into misguided beliefs and behaviours and shape their views in misleading ways[172]. Long-term consequences include increased susceptibility to manipulation, potential radicalisation, and a deepening mistrust in society.

Parents are understandably concerned that children may inadvertently spread misinformation generated by AI systems without knowing it[173]. Misinformation travels faster than factual information. Research has shown that children often share content without verifying its authenticity, driven by the desire to share intriguing information with others. In an experiment conducted by Ali et al. (2021), where children played a simulation game on how news spreads in the real world, one participant remarked, "I just shared what I found interesting and wanted other people to know." This lack of critical engagement highlights the need for comprehensive digital literacy education.

3.3. Promotion of bias and harmful stereotypes

There is growing evidence that GenAI contents (including both text-to-text and text-to-image outputs) can perpetuate and exaggerate societal biases and negative stereotypes, particularly those related to ethnicity, race, gender, and sexual orientation[174][175]. Evaluations of three popular generative AI tools, including Midjourney, Stable Diffusion, and DALL·E 2[176][177], found significant biases against women and African Americans and prejudices in portraying emotions and appearances. For example,

[169] Department for Education. (2023). Generative AI in education Call for Evidence: summary of responses
[170] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models.
[171] UNICEF. (2021). Digital misinformation/disinformation and children.
[172] Torney-Purta, J. V. (2017). The development of political attitudes in children. Routledge.
[173] Analyses were based on Reddit posts and interviews with 7 children and 13 parents in the US. See more: Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications.
[174] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models.
[175] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.
[176] Study of Midjourney, Stable Diffusion, and DALL·E 2 tools. See more: Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in Generative AI.
[177] Although these tools are not chatbots themselves, they can be integrated into conversational AI systems to provide visual responses or image generation capabilities

women were often depicted as younger, more smiling, and happier, while men were portrayed as older, with more neutral or angry expressions. This portrayal risks reinforcing the perception of women as more submissive and less competent than men. The study found that the evident gender and racial biases were even more pronounced than the status quo when compared to labour force statistics or Google images. Independent reviews of generative AI models, including DALL-E and Stable Diffusion[178], have identified similar risks, such as the tendency toward the objectification and sexualisation of women and girls, as well as the reinforcement of gender stereotypes. Although the degree of bias varied across different models, the general direction remained consistent across commercial and open-source AI generators. The ease with which AI-generated images can be created has accelerated the spread of biased content, further entrenching stereotypes and promoting a distorted view of societal norms. Guilbeault et al. (2024) emphasise how the increasing prevalence of image-based content significantly influences gender biases[179].

Exposure to such biased content can have impacts on children's psychological development, identities, and worldviews. There is limited literature on how exposure to such GenAI-induced bias can affect children specifically, but existing studies in child development have demonstrated that exposure to stereotypes at a young age can have long-lasting effects on children's self-esteem and career aspirations, i.e., girls are often discouraged from pursuing STEM subjects, while boys may avoid the arts[180][181]. Additionally, when children encounter AI-generated images that reflect biased representations, they may internalise these portrayals and affect their perceptions of gender roles and racial identities, which is especially concerning during the critical phase of developing their identity and worldview (i.e., middle school age).

Research has shown that children could identify bias in AI-generated images and their potential harms. However, they often struggle to understand the technical mechanisms behind these biases. Explainability (Explainable AI) in educational activities could improve awareness bias among children. Vartiainen et al. (2024)[182] conducted an experiment where children were exposed to AI-generated images and could identify biases, such as those related to gender, appearance, and age, particularly in depictions of occupations such as firefighters. The children recognised the potential harm these biases could cause, such as body appearance pressure, yet they often attributed these biases to human thinking and actions, such as gender-based inequality, rather than the technical mechanisms behind these biases (e.g., how AI systems learn from data and make decisions based on training data sets). In another study, Melsión

[178] Common Sense Media. (2023). Common Sense Media Launches First-Ever AI Ratings System (accessed on 15th August 2024)

[179] Guilbeault, D., Delecourt, S., Hull, T., Desikan, B. S., Chu, M., & Nadler, E. (2024). Online images amplify gender bias. *Nature*, *626*(8001), 1049-1055.

[180] Smith, C. S., & Hung, L. C. (2008). Stereotype threat: Effects on education. *Social Psychology of Education*, *11*, 243-257.

[181] Deemer, E. D., Thoman, D. B., Chase, J. P., & Smith, J. L. (2014). Feeling the threat: Stereotype threat as a contextual barrier to women's science career choice intentions. *Journal of Career Development*, *41*(2), 141-158.

[182] Hands-on workshops with fourth- and seventh-grade students ($N = 209$) in Finland. See more: Vartiainen, H., Kahila, J., Tedre, M., López-Pernas, S., & Pope, N. (2024). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New media & Society*, 14614448241252820.

et al. (2021)[183] used visual explanation tools (i.e. Grad-CAM)[184] as an explainability technique to help children improve their understanding of the concept of bias in terms of gender discrimination.

However, bias in text-based AI outputs can be more subtle and harder to detect than visual biases. Children have expressed concerns about how AI-generated text could introduce bias into their work. Williams-Ceci et al. (2024)[185] show how autocomplete suggestions can subtly shift people's attitudes without being fully aware of the bias and influence embedded in them. In a related study by Higgs & Stornaiuolo (2024)[186], children voiced their worries about this issue, with one student stating, *"Probably [my biggest concern is] the inherent bias in AI. Because using AI for writing can lead to you having bias in your own writing without that intention."* Approximately 20% of surveyed children aged 8–18 stated that they simply copied what GenAI provided without adding their thoughts[187], highlighting the risk of how using AI could subtly influence the attitudes and behaviours of young users, often without their conscious awareness.

3.4. Emotional dependency and mental safety risks

An increasing number of children are using these AI tools as digital companions, for friendships, and even for romantic relationships. These AI-driven tools, which include platforms such as Pi (with 100 million users), SimSimi (with 350 million users), Chai (with 4 million active users), and Replika (with 2.5 million active users), offer accessible, scalable, and stigma-free support for young people. However, research has shown that AI-driven mental health counseling can be effective in reducing symptoms of depression, but does not lead to a significant improvement in overall psychological well-being[188].

However, there are substantial concerns regarding the safety, effectiveness, and ethical implications of using GenAI chatbots for mental health. Before the launch of GenAI chatbots, traditional rule-based chatbots such as Woebot, Wysa, and Tess were employed in mental health services for basic Q&A. Their responses, however, can be less engaging and feel artificial. Though, they are controlled and safeguarded

[183] Study were conducted with 78 children aged 10-14. See more: Melsión, G. I., Torre, I., Vidal, E., & Leite, I. (2021, June). Using explainability to help children understandgender bias in AI. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (pp. 87-99).

[184] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, *128*, 336-359.

[185] Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., Naaman, M., & Tech, C. (2024). Bias in AI Autocomplete Suggestions Leads to Attitude Shift on Societal Issues.

[186] Higgs, J. M., & Stornaiuolo, A. (2024). Being Human in the Age of Generative AI: Young People's Ethical Concerns about Writing and Living with Machines. *Reading Research Quarterly*.

[187] Survey was conducted on 15,830 young people aged 13 to 18 and 1,228 teachers from schools across the UK. See more: National Literacy Trust. (2024). Children, young people and teachers' use of generative AI to support literacy in 2024.

[188] Li, H., Zhang, R., Lee, Y. C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, *6*(1), 236.

through pre-scripted responses and pre-defined rules[189][190]. GenAI chatbots, on the other hand, can offer more personalised responses and dynamic interactions, but their unpredictability and black-box nature raise concerns. Building mostly on deep learning techniques, GenAI responses are unpredictable and, in many cases, unhelpful or even harmful. Concerns have also been raised on how interacting with AI companions chatbots could rewire how children process emotional and social interactions[191]. Research has shown that these AI tools struggle to understand emotional nuances, fail to appropriately detect and handle sensitive topics, such as sexual harassment, and lack the ability to engage in empathetic listening or to interpret subtle emotions accurately[192]. In many cases, these AI companions may fail to recognise distress signals or respond empathetically, potentially exacerbating mental health crises instead of alleviating them.

Using GenAI chatbots as digital companions introduces risks of emotional dependence, negative impact on social skills, and potential harm to the mental safety of children. Children can risk forming unhealthy emotional attachments when they anthropomorphise the chatbots. Young users may increasingly rely on these chatbots for emotional support or social interaction, which could hinder the development of essential social skills and real-life relationships. The allure of frictionless, judgment-free interactions with AI may lead to a preference for these digital relationships over more complex human interactions, potentially degrading the quality of social connections in children's lives and difficulties in forming genuine human relationships.

For example, teenagers in Yu et al. (2024) reported heavy usage of Character.ai, leading to a loss of control and negative impacts on their social lives. One user expressed, "I spent too much time on c.ai. I wanna be able to talk to other people my age, like at my school or something." This suggests that GenAI chatbots are often used to fill a void in personal connections, resulting in unhealthy dependency. Further exemplifying this issue, other users shared similar sentiments on the same subreddit. One user stated, "I feel like a loser right now. I literally had a whole relationship with an AI girl…" Another questioned, "Character.ai is down, how am I supposed to keep my suicidal thinking at bay?" These posts highlight the negative effects of over-reliance on GenAI chatbots on teenagers' mental well-being and stability[193].

[189] Systematic literature review and meta analysis of chatbots on mental health. Identified 12 studies examining the effect of using chatbots on 8 outcomes. See more: Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of medical Internet research*, *22*(7), e16021.

[190] Examination of three chatbots, Woebot, Wysa and Tess, and their role in mental health recovery. Meadows, R., Hine, C., & Suddaby, E. (2020). Conversational agents and the making of mental health recovery. *Digital health*, *6*, 2055207620966170.

[191] Toppo, G. (2024). AI 'Companions' are Patient, Funny, Upbeat — and Probably Rewiring Kids' Brains. (accessed 27th September 2024)

[192] De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: insights into the safety of generative AI. *Journal of Consumer Psychology*, *34*(3), 481-491.

[193] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications. arXiv preprint arXiv:2406.10461.

3.5. Cognitive risks

Early evidence suggests that overreliance on GenAI chatbots can lead to a decline in core cognitive skills, such as critical thinking, analytical abilities, and creativity[194]. The personalised and interactive nature of these AI chatbots can foster a deeper cognitive dependence on the technology, potentially diminishing users' inclination to engage in independent cognitive processes. By offering personalised responses and adaptive conversations, chatbots might inadvertently discourage users from exercising critical thinking and problem-solving skills, as they bypass essential cognitive steps that would normally be involved in these tasks. This phenomenon, known as "cognitive offloading," occurs when users rely on AI to handle tasks that would otherwise require significant mental effort[195]. An overreliance on AI as a primary source of information can weaken children's ability to critically assess information and develop independent research skills, echoing concerns about the shallow processing of information in the digital age.

Instant gratification from interactions with the GenAI chatbots can reduce attention spans in children. Gratification is defined as "expectations about the content or media-related satisfaction to be derived from consumption"[196]. Research has found a strong correlation between the social gratification users experience and their overall satisfaction with AI-powered chatbots[197]. This instant gratification, while satisfying in the short term, can reduce attention spans, particularly in children, and can be linked to excessive screen time and associated attention problems[198]. When children are exposed to environments where immediate feedback or rewards are prevalent, it can condition them to prefer shorter tasks and show less patience for longer, more challenging activities. The constant exposure to rapidly changing content can train the brain to expect high levels of stimulation, making it difficult for children to engage in activities that require prolonged focus, such as reading or completing homework[199]. However, it should also be noted that these effects can differ depending on individual user traits, the type of digital technology used, the context of use, and how these factors interact[200].

---

[194] Dergaa, I., Ben Saad, H., Glenn, J. M., Amamou, B., Ben Aissa, M., Guelmami, N., ... & Chamari, K. (2024). From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. *Frontiers in Psychology*, *15*, 1259845.

[195] León-Domínguez, U. (2024). Potential cognitive risks of generative transformer-based AI chatbots on higher order executive functions. *Neuropsychology*.

[196] Palmgreen, P., Wenner, L. A., & Rayburn, J. D. (1980). Relations between gratifications sought and obtained: A study of television news. *Communication Research*, *7*(2), 161–192.

[197] Xie, C., Wang, Y., & Cheng, Y. (2024). Does artificial intelligence satisfy you? A meta-analysis of user gratification and user satisfaction with AI-powered chatbots. *International Journal of Human–Computer Interaction*, *40*(3), 613-623.

[198] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.

[199] Zimmerman, A., Janhonen, J., & Saadeh, M. (2023). Attention Span and Tech Autonomy as Moral Goods and Societal Necessities. *Digital Society*, *2*(2), 23.

[200] Vedechkina, M., & Borgonovi, F. (2021). A review of evidence on the role of digital technology in shaping attention and cognitive control in children. *Frontiers in Psychology*, *12*, 611155.

3.6. Privacy and security risks

Among different types of privacy, children are most prone to commercial privacy, which relates to the harvesting and use of personal data for business and marketing purposes. Livingstone et al. (2019)[201] identified three types of online privacy relevant to children: (i) **interpersonal privacy**, which deals with how a child's 'data self' is created, accessed, and shared through online social interactions; (ii) **institutional privacy**, which involves how public entities, such as government, educational, and healthcare institutions, collect and manage data about the child; and (iii) **commercial privacy**, which relates to the harvesting and use of personal data for business and marketing purposes. Evidence indicates that *commercial privacy* is the area where children are least prepared to protect themselves. They often struggle to grasp the full complexity of internet data flows and the commercial use of their data. There are also growing concerns about *institutional privacy*, raising questions about informed consent and children's rights.

Children's understanding of privacy is not fully developed, especially at younger ages, making them particularly vulnerable to data collection practices, data misuse and breaches. For instance, children under the age of 7 generally lack an abstract understanding of concepts such as "privacy" and "safety" and while this understanding improves by age 11, they still often struggle to apply these concepts effectively in real-life situations[202]. Additionally, children under 11 are more inclined to share their personal information on websites that feature warnings about age-inappropriate content, as such warnings tend to provoke their curiosity [203][204]. This can lead to children inadvertently sharing personal information, such as health conditions, school details, or family information, during their interactions with chatbots without fully grasping the potential consequences[205][206].

Moreover, design elements such as emotional engagement, human-likeness, personalisation, and self-disclosure in GenAI chatbots can encourage users to disclose more information than necessary, compromising the user's privacy[207]. For example, GenAI chatbots are found to employ "nudging" strategies—subtle prompts that steer conversations in ways that encourage users to share more information than they intended. This is particularly problematic for children, who may not have the cognitive maturity

[201] Livingstone, S., Stoilova, M., & Nandagiri, R. (2019). Children's data and privacy online: growing up in a digital age: an evidence review. London: London School of Economics and Political Science.
[202] Kumar, P., Naik, S. M., Devkar, U. R., Chetty, M., Clegg, T. L., & Vitak, J. (2017). 'No Telling Passcodes Out Because They're Private' Understanding Children's Mental Models of Privacy and Security Online. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-21.
[203] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.
[204] Lwin, M. O., Stanaland, A. J., & Miyazaki, A. D. (2008). Protecting children's privacy online: How parental mediation strategies affect website safeguard effectiveness. *Journal of Retailing*, *84*(2), 205-217.
[205] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts.
[206] Creane, M. W. (2024). Bolts or Brains: How elementary school children conceptualise AI and reason about issues of personal disclosure and privacy. In *School Children and the Challenge of Managing AI Technologies* (pp. 57-68). Routledge.
[207] Experiments were conducted with Blenderbot by Meta and ChatGPT-4 by OpenAI to investigate privacy harms and risks. Gumusel, E., Zhou, K. Z., & Sanfilippo, M. R. (2024). User Privacy Harms and Risks in Conversational AI: A Proposed Framework.

to provide informed consent or recognise when they are being nudged into revealing sensitive details. The friendly and empathetic design of many GenAI chatbots can blur the lines between human and machine interactions, leading children to trust these systems and share sensitive information more readily than they would with a human.

Additionally, the extensive data collection involved in the functioning of GenAI chatbots raises significant concerns poses risks of unauthorised access, data breaches, and misuse, particularly when it involves sensitive information from children. These systems gather vast amounts of personal data, including user demographics, conversation histories, and even browsing behavior, to provide personalised experiences. The extensive data collection involved in these interactions—often done without transparent user consent—leaves the information vulnerable to unauthorised access, data breaches, and potential misuse. For instance, in a recent case, the UK's Information Commissioner's Office (ICO) launched an investigation into a GenAI company's data practices, marking its first action against the use of such technology in relation to data protection[208]. This highlights the regulatory challenges associated with protecting children's data in the rapidly evolving AI landscape. Moreover, current regulations, such as those in the UK, allow individuals to request that their data be deleted from an organisation after a certain period. However, with GenAI chatbots, even if the data is deleted, the underlying algorithms may have already learned from it, making true data deletion nearly impossible. This complicates the "right to be forgotten" and highlights the challenges of ensuring digital privacy for children.

3.7. AI-generated CSAM

Parents have voiced their concerns over the risk of external parties exploiting children's personal information, especially children's faces, to create AI-generated child sexual abuse material (AI CSAM)[209].

AI CSAM has increased in volume, severity and accessibility. The number of pieces of (AI-CSAM) in the UK, though initially small (e.g. accounting for less than 1% of CSAM shared among known child sex abuse networks), has shown a steady increase since August 2022[210] (e.g. there is an increase of 22% in the surveyed period from September 2023 to March 2024)[211]. Similar trends have been observed in the US, where AI-generated CSAM accounts for less than 1% of the reports but increased significantly (e.g. in 2023, the CyberTipline ® received 4,700 reports of CSAM or other sexually exploitative content related to GenAI. The number in practice is likely higher due to unreported issues).[212]. Additionally, AI-generated videos, which were rarely found before, have now been found on dark web forums, indicating the adoption of new deepfake technology among child abuse communities. Analysis of the characteristics

[208] Information Commissioner's Office. (2024). ICO warns organisations must not ignore data protection risks as it concludes Snap My AI chatbot investigation. Information Commissioner's Office.
[209] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications.
[210] Thiel, D., Stroebel, M., & Portnoff, R. (2023). Generative ML and CSAM: Implications and Mitigations.
[211] From April 2023 to March 2024, there were a total of 375 reports containing AI-generated content, in which, 70 reports included criminal AI-generated child sexual abuse materials (CSAM). IWF. (2024). What has changed in AI CSAM landscape?
[212] National Center for Missing and Exploited children. (2023). CyberTipline Report.

of the detected AI-CSAM found: (1) *these images increasingly look real, making them indistinguishable from actual images* (e.g., 82% of images assessed being realistic enough to be assessed as pseudophotographs of children); (2) *there is a surge in CSAM that can be classified as the most severe* (e.g., an increase of 10% in CSAM classified as category A), showing that perpetrators are successfully generating more complex and explicit scenarios; *(3) AI CSAM can be generated to resemble specific children, which further victimises of those children.* There are active communities that share hints and tips on generating CSAM from open-source models. The users can download the models, access them offline, and generate CSAM offline, further evading detection by law enforcement[213].

However, AI CSAM is not only shared and available on dark web forums, but also penetrate clear webs, especially social media platforms. AI bots on social media platforms (i.e., Telegram) have been found to create and widely share deepfakes of inappropriate images of children without their knowledge or consent[214]. With the open-source models and ease of generating CSAM from text-to-image, or text-to-video, AI CSAM can be created without sophisticated technical knowledge. Children can inadvertently share their data through app-harvesting photos such as FaceApp, FaceTune, Magic AI Avatars, and Lena[215] without being aware that those photos could be used for malicious purposes, as warned by FBI[216]. Active communities, including Reddit users and dark web users, have shown support for creating and sharing deepfake artifacts that are related to minors[217][218]. Although the generation, possession and distribution of deepfake CSAM is illegal in the UK under the Online Safety Act[219], tutorials and guides on how to create realistic AI-generated CSAM remain legal and are still widely shared among members of these communities[220].

The generation, sharing, and distribution of AI CSAM can have devastating effects not only on victims and survivors and their families but also on society. *Firstly*, targeted deepfake CSAM can cause long-term psychological harm such as anxiety, panic, depression, and PTSD, privacy violations, cyberbullying; reputation damage, identity theft, blackmail, financial issues through extortion; loss of trust in technology; and, in extreme cases, loss of life[221]. The Internet Watch Foundation (IWF), in partnership with the UK Safer Internet Centre, reported an eightfold increase in financially-motivated sexual extortion (sometimes referred to as 'sextortion' cases over 2022–2023 (21 vs. 176 cases). Teenage boys, appearing in 91% of the reports, are being deliberately targeted, often blackmailed for money[222]. *Secondly*, the

[213] IWF. (2023). How AI is being abused to create child sexual abuse imagery.
[214] Wired. (2020). Telegram Still Hasn't Removed an AI Bot That's Abusing Women.
[215] Heikkilä, M. (2022). The viral AI avatar app Lensa undressed me—without my consent. *MIT Technology Review*.
[216] FBI. (2023). Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes.
[217] Study of 13293 comments on Reddit. Eelmaa, S. (2022). Sexualization of children in Deepfakes and hentai. *Trames*, *26*(2), 229-248.
[218] Study of 6,638 posts and 86,425 comments on Reddit from 2018 to 2021. Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M. E., & Sasahara, K. (2022, April). Are deepfakes concerning? analyzing conversations of deepfakes on reddit and exploring societal implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).
[219] Cooney, C. (2024). Creating sexually explicit deepfakes to become a criminal offence. BBC News.
[220] IWF. (2024). What has changed in AI CSAM landscape?
[221] Romero Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, 1-30.
[222] UK Safer Internet center. (2024). Sextortion report of boys increase as IWF announces record amount of cases.

consumption of AI CSAM can desensitise users and escalate fantasy, thus increasing the risks of actual child abuse and violence[223][224]. *Thirdly*, the sheer volume of AI CSAM complicates the efforts of law enforcement, NGOs, and tech companies in identifying actual victims and requires significant resources to remove CSAM.

IV. REGULATING GENAI FOR CHILDREN

4.1. Overview of current frameworks on GenAI and children and practical implementation

The UK's principles-based approach to regulating AI, including GenAI, demonstrates its flexibility and agility in responding to emerging technologies: When discussing the regulation of GenAI for children, it is essential to first consider the broader framework within which the UK operates. The UK adopts a distinct, principles-based and pro-innovation approach to regulating AI through existing sector regulators. It is guided by five key principles: (1) safety, security, and robustness; (2) transparency and explainability; (3) fairness; (4) accountability and governance; and (5) contestability and redress[225][226]. In contrast, the Biden-Harris administration in the US has taken a more directive route, providing specific guidance for federal agencies through Executive Order, while the European Union has implemented a more prescriptive, risk-based model under the European AI Act, which categorizes AI uses into four risk levels and imposes specific requirements for general-purpose AI models.

However, critics suggest that the US approach, which mandates actions for federal agencies and depends on voluntary commitments from leading developers, may not have strong enough enforcement measures[227]. Meanwhile, the EU's approach has faced criticism for being prescriptive, rigid, and top-down, raising concerns about inconsistent implementation across its member states[228]. The EU's risk-based model, which focuses on minimising risks in specific contexts, has been criticised for focusing more on traditional AI rather than addressing the unique challenges posed by GenAI because: 1) it is impossible to specify all potential downstream contexts of use and stakeholders, and while the risks are significant, they are difficult to predict and assess; and 2) the ethical principles guiding regulation of "high-risk applications" in the AI Act do not sufficiently extend to evaluating the unique risks posed by GenAI[229]. The EU's emphasis on regulating AI models rather than addressing societal risks from AI-driven content may leave gaps in managing the immediate dangers posed by misinformation and harmful content on digital platforms,

[223] Christensen, L. S., Moritz, D., & Pearson, A. (2021). Psychological perspectives of virtual child sexual abuse material. *Sexuality & Culture*, *25*(4), 1353-1365.
[224] Protect Children. (2021). CSAM users in the darkweb. Protecting children through prevention.
[225] House of Commons. (2024). Governance of artificial intelligence (AI). Third Report of Session 2023–24.
[226] Department for Science, Innovation and Technology. (2024). A pro-innovation approach to AI regulation. Government response to consultation.
[227] House of Commons. (2024). Governance of artificial intelligence (AI). Third Report of Session 2023–24.
[228] House of Commons. (2024). Governance of artificial intelligence (AI). Third Report of Session 2023–24.
[229] Elgesem, D. (2023). The AI Act and the Risks Posed by Generative AI Models. In *NAIS*. The 2023 symposium of the Norwegian AI Society, June 14-15, 2023, Bergen, Norway.

highlighting a potential misalignment between the AI Act's focus and the urgent need to address societal risks[230].

In contrast, the UK's principles-based approach offers more flexibility by allowing sector-specific regulators to address AI risks adaptively, especially in the context of children exposed to downstream GenAI applications. While the UK's model is more agile, it may still need to strengthen its regulatory framework to manage emerging issues such as misinformation and AI-generated content[231].. However, despite its flexibility, the principles-based approach may face difficulties in implementation and operationalisation in practice[232] and present challenges for effective regulation due to varied interpretations[233]. We will discuss this in detail in the later section regarding children specifically.

However, it is worth acknowledging that there has been a small but growing number of frameworks specifically focusing on children[234][235]. In the UK, this issue is addressed through laws, regulations and frameworks such as the Online Safety Act, the UK Age-Appropriate Design Code, the General Data Protection Regulation (GDPR), and the 5Rights Foundation's "Playful by Design" initiative. Internationally, notable frameworks include the UN Committee on the Rights of the Child's General Comment 25 (UNCRC), the IEEE's Standard for an Age-Appropriate Digital Services Framework, the World Economic Forum's Artificial Intelligence for Children toolkit, and UNICEF's Policy Guidance on AI for Children. A brief explanation and critiques of each of these frameworks is provided in the following tables. Review of different frameworks are documented in research of Wang et al. (2022)[236], Mahomed et al., (2023a)[237], Mahomed et al., (2023b)[238], and Caivano et al., (2024)[239]. The reviews show that **navigating regulatory and governance requirements remains challenging, and there is a lack of clear guidance on specific actions to take, makes it challenging for designers and practitioners to effectively develop concrete design standards.**

---

[230] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).
[231] House of Commons. (2024). Governance of artificial intelligence (AI). Third Report of Session 2023–24.
[232] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).
[233] Wood, S. (2024). Impact of regulation on children's digital lives.
[234] Mahomed, S., Aitken, M., Atabey, A., Wong, J., & Briggs, M. (2023). AI, Children's Rights, & Wellbeing: Transnational Frameworks.The Alan Turing Institute.
[235] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).
[236] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).
[237] Mahomed, S., Aitken, M., Atabey, A., Wong, J., & Briggs, M. (2023a). AI, Children's Rights, & Wellbeing: Transnational Frameworks.The Alan Turing Institute.
[238] Mahomed, S., Briggs, M., Wong, J., & Aitken, M. (2023b). Navigating Children's Rights and AI in the UK: A roadmap through uncertain territory.
[239] Caivano, D., Nieto, B. F., Gigante, D., Ragone, A., & Tibidò, S. (2024, September). Ensuring Child Rights in the Age of AI: A Multidimensional Analysis of Existing Frameworks. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 76-83).

Table : Overview of relevant frameworks on GenAI and children

Organisation	Title	Legal status	Brief explanation
UK Government	The Online Safety Act (OSA)	Legally-binding	**The Online Safety Act (OSA)** passed into law in 2023. The OSA covers online safety for all users and contains specific obligations related to children. The overall approach of the legislation is **design-focused and risk-based.** The safety duties contained in the OSA are to be considered against other rights, such as freedom of expression. The duties are also framed as a requirement to use proportionate measures. The OSA states that 'service providers which implement measures recommended to them in the children's codes will be treated as complying with the relevant duty or duties to which those measures relate'. However, it is criticised that while the Act includes certain platform design regulations, it does not fully commit to this approach and **struggles with ensuring proportionality and accountability**. The OSA's focus on traditional content controls may limit its effectiveness in tackling emerging online safety challenges as digital services continue to evolve[240].
UK Government (Information Commissioner's Office - ICO)	The UK Age Appropriate Design Code (AADC)[241]	Legally-binding	The Code took effect in 2021. It is a set of 15 standards that online services should meet to protect children's privacy. The document sets out the standards expected of those responsible for designing, developing or providing online services that are likely to be accessed by children. **As a regulatory measure, it focuses on protecting children from harms associated with the use of their data and safeguarding their privacy. However, it does not specifically address issues related to content or general safety**[242]. **The AADC does cover areas where data and content intersect, such as when profiling is used to target content at children**. Importantly, it applies to all online services "likely to be accessed by children," not just those explicitly designed for them.
EU	General Data Protection Regulation (GDPR)	Legally-binding	The General Data Protection Regulation (GDPR) was fully implemented in 2018 across both the EU and the UK. Since Brexit, the UK has maintained GDPR as a separate law, supported by the **UK Data Protection Act 2018.** The GDPR includes several provisions relevant to children's online safety, emphasising the need for specific protections for children's personal data and their rights in relation to data processing. When online services rely on consent as a lawful basis for processing children's personal data, parental consent is required for children under 16. In the UK, the age is set at 13.

[240] Nash, V., & Felton, L. (2024). Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation. *Policy & Internet*.
[241] ICO. (2020). Age appropriate design code.
[242] Wood, S. (2024). Impact of regulation on children's digital lives.

European Parliament and Council of the European Union	EU Audiovisual Media Services Directive (AVMSD)	Legally-binding	The Directive primarily focuses on regulating content provided by broadcasters, on-demand services, and video-sharing platforms (VSPs), requiring measures to protect children 'against content which may impair their physical, mental or moral development'. These measures include establishing terms and conditions for users, implementing age verification systems, and creating mechanisms for reporting harmful or illegal material. While the AVMSD **offers a more limited set of safety measures compared to the Online Safety Act (OSA)**, it might play a role in shaping the design and implementation of age verification systems on certain platforms[243].
International frameworks			
UN	The UN Committee on the Rights of the Child's endorsement of General Comment No. 25 (UNCRC)[244]	Voluntary guidance	This guidance outlines four key principles for protecting children's rights in the digital environment: (1) **non-discrimination**, ensuring equal access to meaningful digital experiences for all children; (2) **best interests of the child**, prioritising children's well-being in all digital-related decisions and actions; (3) **right to life and development**, protecting children from digital threats such as violent content, harassment, and exploitation while emphasising technology's role in their early development and educating caregivers on safe technology usage; and (4) **respect for the child's views**, encouraging children's expression on digital platforms, integrating their feedback into policies, and ensuring their privacy and freedom of thought are respected by service providers. Though not legally binding, General Comment No. 25 is an influential international framework guiding member states on how to create legislation and regulations that protect children's rights online. It marked a significant milestone in safeguarding children's rights in the digital age, addressing the full spectrum of children's rights, including but not limited to data protection, privacy, and safety, offering a holistic view that aligns with the evolving nature of children's digital lives.
IEEE	Standard for an Age Appropriate Digital Services Framework Based on the 5 Rights	Voluntary guidance	The IEEE Framework is structured around a series of processes that outline key concepts for creating age-appropriate AI products and services. Each process details its purpose, outcomes, tasks, and inputs/outputs, **making the framework more accessible for practitioners to implement.** However, one potential limitation is that **some of the suggested tasks, such as conducting impact assessments on children's rights, are still vague and lack specific guidance**, which may hinder practical application for teams and stakeholders[246].

[243] Wood, S. (2024). Impact of regulation on children's digital lives.

[244] OHCHR. (2020). General comment No. 25 (2021) on children's rights in relation to the digital environment.

[246] Caivano, D., Nieto, B. F., Gigante, D., Ragone, A., & Tibidò, S. (2024, September). Ensuring Child Rights in the Age of AI: A Multidimensional Analysis of Existing Frameworks. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 76-83).

	Principles for Children[245]		
World Economic Forum (WEF)	Artificial Intelligence for Children - Toolkit[247]	Voluntary guidance	The Toolkit offers tailored guidance for various stakeholders involved in AI development for children, including a checklist for corporate decision-makers, guidelines for product teams, and a guide for parents or guardians. It follows the FIRST principles: **Fair, Inclusive, Responsible, Safe, and Transparent**. However, a potential limitation is that some recommendations, such as incorporating feedback from users and guardians, are still generic and lack detailed methodologies, which may reduce their practical applicability[248].
UNICEF	Policy guidance on AI for children[249]	Voluntary guidance	The framework outlines nine key requirements for child-centered AI and provides recommendations for their implementation. It includes complementary resources, practical tools, and case studies to illustrate real-world applications, enhancing its usability for applying children's rights throughout the AI system lifecycle. However, as a policy-focused document, it lacks specific technical guidance on how to fulfill these requirements, which may limit its direct practical use for developers[250].

Independent evaluation on changes made by companies following the implementation of regulatory frameworks such as the Age Appropriate Design Code (AADC), Digital Services Act (DSA), and Online Safety Act (OSA) show positive developments towards improving children's privacy and safety online in the UK. However, significant challenges remain, including an over-reliance on parental controls, insufficient clarity on addressing harmful content and the impact of GenAI tools, and a transparency gap that hinders effective monitoring and enforcement.

***Firstly,* Wood (2024)[251]** examined changes made by Meta (Instagram, Facebook, Messenger, and Quest), Google (including Youtube), TikTok, and Snapchat in response to regulations and legislation from 2017 to 2024, recording 128 changes in total, with a peak of 42 changes in 2021 when the Age Appropriate Design Code (AADC) took effect. The term "change" refers to any modification in the design, operation, or governance of online services aimed at improving children's privacy and safety. Even though changes made by companies do not address specifically any particular legislation, the research concludes that regulation is prompting companies to improve child privacy and safety online, including **defaulting social media accounts to private settings, adjustments to recommender systems, and restrictions on targeted advertising to children**. However, the study also highlighted concerns about **over-reliance on parental**

[245] CEN IEEE and CENELEC. (2023). Age appropriate digital services framework.
[247] World Economic Forum. (2022). Artificial Intelligence for children. TOOLKIT.
[248] Caivano, D., Nieto, B. F., Gigante, D., Ragone, A., & Tibidò, S. (2024, September). Ensuring Child Rights in the Age of AI: A Multidimensional Analysis of Existing Frameworks. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 76-83).
[249] UNICEF. (2021). Policy guidance on AI for children.
[250] Caivano, D., Nieto, B. F., Gigante, D., Ragone, A., & Tibidò, S. (2024, September). Ensuring Child Rights in the Age of AI: A Multidimensional Analysis of Existing Frameworks. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 76-83).
[251] Wood, S. (2024). Impact of regulation on children's digital lives.

controls, which may not be widely used or effective, and could limit child autonomy. Additionally, the report raises concerns that modifications to age assurance and recommender systems could affect other children's rights, such as freedom of expression and non-discrimination.

***Secondly,* Mootz and Blocker (2024)**[252] analyse the AADC and its impact on creating safer, more age-appropriate online experiences for children and teens. The report identified 91 changes made by platforms, including YouTube, TikTok, Snapchat, Instagram, Marketplace, and Google Search, using publicly available announcements from tech companies between May 2018 and September 2023. These changes cover four categories: age-appropriate design, time management, privacy/security and data management, and youth safety and well-being. The report concluded that companies have made key improvements, including **stricter default privacy settings, reduced content profiling, and fewer notifications.** However, **not all platforms fully addressed all fifteen standards of the Code, and each platform progressed in different ways.** The AADC has also expanded the scope of protections, extending to adolescents up to age 17, which was previously limited to age 13. The non-prescriptive nature of the AADC allows companies **flexibility** in their responses, but this also introduces **ambiguity**, making it harder to **monitor compliance** as companies may implement changes in different ways. Future research should investigate how companies are interpreting and applying AADC standards and identify areas where further guidance or specificity may be needed to ensure comprehensive and consistent progress.

However, in both evaluation reports, it remains unclear how many of the changes made by companies are specifically related to GenAI tools and services embedded within their platforms. The reports emphasise the use of automation and AI to remove intimate child abuse materials (categorised as *Primary Priority Content*). However, there is a lack of clarity on the measures in place and the effectiveness of efforts to mitigate children's exposure to other types of *Primary Priority Content* (e.g., self-harm, eating disorders, or suicide) and *Priority Content* (including bullying, abusive or hateful language), and risks from children's interactions with GenAI chatbots (e.g., nudging for more personal information or potential mental health safety issues). However, it is also worth noting that in terms of technical feasibility, while there have been advancements in automated classification and identification of CSAM, challenges remain in effectively detecting other harmful materials, including hateful content, disinformation, and misinformation. These ongoing difficulties highlight the need for a more holistic approach to content moderation and child safety on digital platforms. Additionally, the reports identified a transparency gap. Both evaluations are based on company announcements, which may not capture all the changes made by the platforms. For instance, Meta is currently the only company to have published a timeline of updates related to child privacy and safety, though this timeline is neither comprehensive nor fully inclusive[253]. This highlights the need for a consistent and regulated approach to ensure transparency, prevent selective presentation of information, and improve accessibility of data for researchers.

4.2. Gaps/challenges in regulating GenAI for children

REGULATORY CHALLENGES:

[252] Mootz, J. & Blocker, K. (2024). UK Age-appropriate Design Code. Impact Assessment.
[253] Wood, S. (2024). Impact of regulation on children's digital lives.

While a principles-based approach to AI regulation, including GenAI, offers agility in adapting to new developments, it remains abstract and can present challenges for effective operationalisation: The abstract nature of principles means they can be interpreted in multiple ways, which can lead to inconsistencies or misinterpretation or even misuse. For instance, a recent report by Livingstone et al. (2024)[254] highlights how, in the context of the digital environment, the "best interests" principle is frequently misunderstood or even misused. The report argues that in most cases, invoking "best interests" is unnecessary, and instead, the emphasis should be on upholding all of children's rights under the UN Convention on the Rights of the Child (UNCRC). A "best interests" determination should only be made when rights are in conflict or third-party claims jeopardise children's rights, providing a standard of conduct expected from digital service providers.

Although various general guidelines and regulations on AI have been established, there is limited focus on the unique challenges and requirements of children. These frameworks frequently fail to consider the diverse developmental stages of childhood, overlooking the specific needs and capabilities of children at different ages[255]. A UNICEF review of 20 national AI strategies in 2020 revealed that very little focus has been placed on safeguarding children's rights, highlighting a significant gap in addressing child-specific concerns in the development and regulation of AI technologies[256]. Few AI frameworks emphasise the importance of giving extra consideration to systems that children may access, particularly the need for child-focused impact assessments[257]. Children's intellectual and emotional development varies, highlighting a need for more nuanced and comprehensive approaches that go beyond simple age-based categories. Most of them adopt a policy-driven approach and often fail to address children's rights from a technical perspective. As a result, they frequently struggle to provide clear, actionable guidance for practitioners on how to integrate and safeguard children's rights throughout the AI system life cycle[258]. This lack of targeted attention highlights a critical shortfall in addressing the specific needs and vulnerabilities of children within AI frameworks.

There is currently a lack of meaningful child participation and parents/teachers participation in shaping AI policy and practice, beyond their roles as users[259]. While many stakeholders express a desire to involve children in decision-making processes related to AI, they often lack the necessary expertise, skills, or resources to do so effectively[260]. Additionally, ethical AI principles rarely address the critical role

---

[254] Livingstone, S., Cantwell, N., Özkul, D., Shekhawat, G., & Kidron, B. (2024). The best interests of the child in the digital environment.

[255] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).

[256] Penagos, M., Kassir, S., & Vosloo, S. (2020). National AI strategies and children. *UNICEF Office of Global Insight and Policy*.

[257] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, *6*(3), 265-270.

[258] Caivano, D., Nieto, B. F., Gigante, D., Ragone, A., & Tibidò, S. (2024, September). Ensuring Child Rights in the Age of AI: A Multidimensional Analysis of Existing Frameworks. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 76-83).

[259] La Fors, K. (2024). Toward children-centric AI: a case for a growth model in children-AI interactions. *AI & society*, *39*(3), 1303-1315.

[260] Mahomed, S., Briggs, M., Wong, J., & Aitken, M. (2023). Navigating Children's Rights and AI in the UK: A roadmap through uncertain territory, 26 September 2023, PREPRINT (Version 1) available at Research Square

of guardians, who make decisions on behalf of children[261]. While guardians play a crucial role, children often possess deeper understanding of AI than their parents or teachers, as shown in earlier sessions. This calls for a shift from a guardian-led approach to a child-centred one that fosters autonomy and resilience.

DESIGN/TECHNICAL CHALLENGES:

Research on design principles for AI systems tailored to children is still in its early stages[262]: The White Paper preceding the Online Safety Act (OSA) introduced a statutory duty of care, emphasising the need for preventative, risk-based, and outcome-focused approaches to digital safety. It highlighted that online platforms do not simply host content; they actively shape user interactions through their design choices[263]. However, this has not yet translated into **comprehensive, child-specific design principles** in AI systems, leaving a gap in fully safeguarding children's digital well-being[264]. While the Human-Computer Interaction (HCI) community has made considerable progress in understanding how children of various ages and backgrounds perceive AI technologies, there is still limited discussion on how to ***systematically design AI systems for children*** (e.g. respecting children's best interests and age-specific needs)[265]. Much of the existing research focuses on designing with and for children in areas such as education, online safety, and entertainment. Recent discussions on age-appropriate design codes have further drawn attention to the unique requirements of children in the digital space. However, there remains a gap in creating frameworks that guide the design of AI systems with a more holistic and child-centred approach.

Technical limitations posing challenges in ensuring the safety of GenAI chatbots: It remains particularly challenging to evaluate the safety of GenAI chatbots. Existing evaluation metrics and safety benchmarks for generative chatbots, such as BLEU and ROUGE, primarily measure linguistic accuracy but fail to capture nuances such as conversational flow, coherence, and emotional impact, thus, often fail to align with human judgment[266]. Additionally, there is a lack of child-centered evaluation[267]. Current AI systems for children rely primarily on quantitative technical evaluations such as accuracy and precision, which may overlook human-centred factors such as children's emotional well-being and long-term development. A more balanced approach that includes both technical and human-centred assessments is necessary. The lack of transparency from companies regarding the performance of their models further complicates the issue, particularly since many downstream applications are developed using these

---

[261] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).

[262] UNICEF. (2020). What do national AI strategies say about children?

[263] Nash, V., & Felton, L. (2024). Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation. *Policy & Internet*.

[264] Reich, S. M., Starks, A., & Su, Z. (2024). One (Adult) Size Does Not Fit All: The Importance of Development in Digital Design and Utilization. *Youth Wellbeing in a Technology Rich World*.

[265] Wang, G., Sun, K., Atabey, A., Pothong, K., Lin, G. C., Zhao, J., & Yip, J. (2023, April). Child-Centred AI Design: Definition, Operation, and Considerations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

[266] Chua, J., Li, Y., Yang, S., Wang, C., & Yao, L. (2024). AI Safety in Generative AI Large Language Models: A Survey. *arXiv preprint arXiv:2407.18369*.

[267] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022, April). Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).

foundational models[268]. This opacity makes it difficult for developers and regulators to fully assess the risks or ethical implications of how these models are applied. Additionally, the interpretability of generative models, particularly deep learning-based ones, is problematic, as they are often viewed as "black boxes" with opaque decision-making processes. Developing more accurate and reliable metrics for assessing these chatbots and improving the interpretability of these models remains an open research challenge[269].

Improving chatbot safety is still a challenging task. Despite efforts to improve chatbot safety, including removing contaminated information and implementing safety layers[270], these models can still produce toxic responses. Si et al. (2022)[271] found that queries related to sensitive topics such as race, those structured as interrogatives, and even generic queries can trigger toxic responses from AI models, highlighting the unpredictability of content generation. Additionally, the lack of universally agreed-upon definitions and categories for toxic content complicates the task of filtering and managing these responses, as social and cultural differences can influence what is considered harmful[272][273].

The lack of transparency in the release of safety evaluations of GenAI models can hinder effective oversight of GenAI chatbot performance, potentially resulting in the deployment of unsafe models. For example, Eiras et al. (2024)[274] found that the majority of evaluated LLMs have their safety evaluation code and data either fully closed or only semi-open. At the same time, researchers have advocated for comprehensive pre-release audits of models, datasets, and research artefacts[275]. However, it is important to acknowledge that while evaluation methods and frameworks (such as HELM and Big-Bench for task evaluation, Chatbot Arena for crowd-sourced model comparisons, and red teaming for exploratory evaluation) offer valuable insights, they provide only a partial view of how models will perform in real-world scenarios. Therefore, ongoing monitoring and evaluation are essential to ensure the safety and reliability of chatbots.

SAFETY MEASURES CHALLENGES:

---

[268] Eiras, F., Petrov, A., Vidgen, B., Schroeder, C., Pizzati, F., Elkins, K., ... & Foerster, J. (2024). Risks and Opportunities of Open-Source Generative AI. *arXiv preprint arXiv:2405.08597*.

[269] Khennouche, F., Elmir, Y., Himeur, Y., Djebari, N., & Amira, A. (2024). Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs. *Expert Systems with Applications*, *246*, 123224.

[270] Xu, J., Ju, D., Li, M., Boureau, Y. L., Weston, J., & Dinan, E. (2020). Recipes for safety in open-domain chatbots.

[271] Si, W. M., Backes, M., Blackburn, J., De Cristofaro, E., Stringhini, G., Zannettou, S., & Zhang, Y. (2022, November). Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2659-2673).

[272] Weng, L. (2021). Reducing toxicity in language models.

[273] Study of a sample of 252 human raters stratified by gender, age group, race/ethnicity group, and locale. See more: Homan, C. M., Serapio-Garcia, G., Aroyo, L., Diaz, M., Parrish, A., Prabhakaran, V., ... & Wang, D. (2023). Intersectionality in conversational AI safety: How Bayesian multilevel models help understand diverse perceptions of safety.

[274] Eiras, F., Petrov, A., Vidgen, B., Schroeder, C., Pizzati, F., Elkins, K., ... & Foerster, J. (2024). Risks and Opportunities of Open-Source Generative AI.

[275] Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). Assessing language model deployment with risk cards.

There is a lack of measures and strategies for managing and monitoring children's engagement with GenAI tools. Some platforms, such as Character.ai, do not have age restrictions in place, and the parental controls integrated into smartphones offer limited capabilities to oversee the use of third-party apps[276]. In the meantime, children can engage with and interact with GenAI tools that might not be made specifically *for* children. Therefore, these independent interactions of children with technology highlight the importance of safe and responsible design, whether with or without adult supervision[277].

The effectiveness of tools such as parental controls and age verification systems remains uncertain and requires further research. The Online Safety Act (OSA) introduces age verification or estimation requirements to protect children from harmful content online, but the implementation raises challenges. Age assurance mechanisms such as hard identifiers or biometrics (i.e., upload their ID, record a video selfie or ask mutual friends to verify their age) are currently being used by companies to enforce compliance. However, Wood (2024)[278] raise concerns that the mechanism could lead to discrimination, limit access to empowering information, or create privacy risks. Further research is needed to explore how age assurance is managed in diverse domestic contexts, particularly for children with disabilities, refugee children without government IDs, or those whose parents are in conflict over their digital activities. Additional challenges include determining when and how to assess age, the accuracy of verification tools (i.e. age estimation), and balancing privacy with security[279]. It is important to acknowledge that age assurance is not a single solution to children's online safety. Ofcom research reveals that many children, especially those aged 8 to 17, sign up for social media with false birthdates, using adult profiles[280].

Additionally, concerns remain that over-reliance on parental control tools might create a false sense of security, lead to excessive surveillance, impact children's independence, and shift responsibility from companies onto parents[281]. Stoilova et al. (2024)[282] shows that parental controls use depending on factors such as the age of the parents and children, their digital skills, parental involvement, and the motivation to reduce exposure to online risk. The outcomes of using parental controls are mixed, offering both benefits and drawbacks, while sometimes limiting or having no impact at all. The review found little support for parental controls as a stand-alone strategy but noted that parents valued them when integrated into broader approaches to parental mediation and parent-child relationships. It is important to recognise that some parental control tools respect children's autonomy and privacy. For instance, Snap's Family Centre allows parents to see who their child has communicated with over the past seven days, though not the content of the conversations, and only with the child's consent. Parental tools are becoming more sophisticated,

[276] See reference 37
[277] Kurian, N. (2024). 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap'in large language models. *Learning, Media and Technology*, 1-14.
[278] Wood, S. (2024). Impact of regulation on children's digital lives.
[279] Nash, V., & Felton, L. (2024). Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation. *Policy & Internet*.
[280] Ofcom. (2024). A third of children have false social media age of 18+.
[281] Wood, S. (2024). Impact of regulation on children's digital lives.
[282] Stoilova, M., Bulger, M., & Livingstone, S. (2024). Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use. *Journal of Children and Media*, *18*(1), 29-49.

evolving from traditional filtering or restrictive models toward more collaborative approaches that enable dialogue between parents and children when access or feature requests arise.

COLLABORATION CHALLENGES:

Lack of collaboration between AI, law, and psychology hinders the development of child-centric principles, calling for more integrated approaches[283]. Each of these fields contributes critical insights that are essential for developing effective guidelines: AI expertise is necessary for understanding the technical capacities and limitations of chatbots, legal insights ensure that regulations protect children's rights and privacy, and psychological knowledge helps tailor these technologies to the developmental needs of young users.

V. POLICY PROPOSALS

This section is a collection of proposals and suggestions from different perspectives: developers, academia, civil society and community (e.g., parents and teachers). Overall, regulating GenAI for children requires a careful balance between protecting children's rights and safeguarding them from harm, while ensuring their privacy, freedom of expression, and protection from discrimination. Regulating GenAI is not only technical solutions but also a socio-technical approach, requiring sociotechnical, geopolitical, and political-economic considerations[284]. Regulatory approaches must be cross-domain, adaptable, and scalable, with the support of independent research, to effectively navigate these complexities and ensure comprehensive child protection in AI environments[285].

5.1. For GenAI developers and deployers:

Transparency obligations: Firstly, GenAI developers and deployers should report on the provenance and curation of training data, performance metrics, and any incidents related to harmful content, along with the mitigation measures (e.g. "model cards")[286][287]. Both developers and third parties should rigorously test these models before deployment to evaluate performance, biases, alignment, and potential risks that may affect children. While "model cards" and training data disclosures are a necessary starting point for transparency, continuous monitoring and auditing the models is required throughout their lifecycle, with changes or issues reported transparently[288][289].

283

284 Leslie, D., & Perini, A. M. (2024). Future Shock: Generative AI and the international AI policy and governance crisis. *Harvard Data Science Review* (special issue 5)

285 Eben, M., Erickson, K., Kretschmer, M., Cifrodelli, G., Li, Z., Luca, S., ... & Schlesinger, P. (2023). Priorities for generative AI regulation in the UK: CREATe response to the Digital Regulation Cooperation Forum (DRCF).

286 Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).

287 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

288 G'sell, F. (2024). Regulating under Uncertainty: Governance Options for Generative AI. *Available at SSRN*.

289 Liang, W., Rajani, N., Yang, X., Ozoani, E., Wu, E., Chen, Y., ... & Zou, J. (2024). What's documented in AI? Systematic Analysis of 32K AI Model Cards.

Transparency also extends to ensuring that information is both understandable and accessible to all users, particularly when systems affect children. This requires providing clear and actionable insights so that parents, caretakers, and educators can comprehend how AI impact on children and what actions can be taken to create safer environment[290]. The World Economic Forum (WEF)[291] introduced **AI labelling systems** that offer key details such as the product's intended age range, accessibility features, sensor use (e.g., cameras, microphones), network capabilities, AI interactions (e.g., facial, voice, emotion recognition), and how user data is collected or shared (e.g., specific data collection practices such as location tracking, behavioral data analysis, or usage patterns). By indicating whether data is shared with third parties or used for targeted recommendations or advertising, these labeling systems aim to enhance transparency and build trust among child and youth users, as well as their parents and guardians. Yu et al. (2024)[292] further recommended that disclosures can incorporate risks and protections associated with the AI products, along with with users' perceptions of risks. By incorporating user feedback on AI safety and privacy, companies can improve the design and presentation of these disclosures, making them more accessible and relevant. This feedback can also guide policymakers in crafting more effective, user-centered regulations.

Independent auditing & bias mitigation: Developers should prioritise proactive audits of training datasets, specifically to address potential biases that may impact children. Given that children are especially vulnerable to stereotypes and misrepresentations, auditing should ensure that protected groups, including gender, race, and age-related characteristics, are accurately represented. For AI systems intended for young users, it is particularly important to scrutinise the nature of the data sources (e.g., curated educational content versus unmoderated social media data) to avoid perpetuating harmful biases. Additionally, incorporating synthetic data alongside real-world data can help counteract historical or societal biases, fostering a more inclusive and fair AI experience for children[293].

Accessibility by researchers: Data relevant to child safety research should be openly accessible to researchers, and companies should be obliged to comply. For example, companies should provide a centralised web portal that allows researchers and stakeholders to monitor updates to child privacy and safety protocols, with changes clearly listed by date and regions/areas those changes applied to. These updates should also be accessible via an API and in machine-readable formats[294].

Child-centered approach/Child-centered AI: A participatory and inclusive design approach should be adopted to actively involve stakeholders such as parents, schools, teachers, practitioners, and especially children from diverse backgrounds[295]. Participatory methods include activities such as focus groups with parents, collaboration with educators, and consulting child psychologists to ensure AI tools are developmentally appropriate and regulations take into accounts perspectives of children. However, a

[290] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, *6*(3), 265-270.
[291] World Economic Forum. (2022). AI Labelling system (accessed 27th September 2024)
[292] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications. *arXiv preprint arXiv:2406.10461*.
[293] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).
[294] Wood, S. (2024). Impact of regulation on children's digital lives.
[295] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, *6*(3), 265-270.

genuine child-centred AI must move beyond superficial features (e.g., child-friendly voices or basic content restrictions) and instead, focus on human factors and designing AI systems with children's best interests in mind, treating them in respectful, fair, and autonomy-supportive ways[296]. Children should be at the forefront of design and implementation, wherein their needs, experiences, and perspectives are reflected and considered. Involving children directly in the design process also serves as a means to empower children and support their autonomy and resilience.

5.2. For regulators

Incorporating children's voices: Similar to participatory design that actively involves children and relevant stakeholders in creating AI systems, regulatory decision-making processes should also integrate children's perspectives at various stages. Further research is essential to gather and examine evidence reflecting children's voices and online experiences to inform policy. For instance, studies could investigate how children interact with default privacy settings on social media platforms, how design changes made by companies impact children's experiences online, or whether they choose non-personalised feeds and how these choices affect their online experiences[297].

Practical guidelines: Lack of practical guidance and resources is highlighted as one of the main obstacles that hinder designers and developers from effectively translating ethical AI principles for children[298]. To address this, regulators should publish clear expectations for good practices and provide guidance on how platforms should document and record changes related to child safety. Additionally, regulators should explore the development of a transparent "child online safety tracking database," where data on child safety updates can be systematically recorded and logged for public access[299].

Industry standards/best practices/multidisciplinary collaboration: A bottom-up approach which includes community building among developers, designers, practitioners, child protection specialists, AI ethicists, and policy-makers, could be formed to not only generate practical resources/best practices but also promote a culture of accountability and continuous improvement[300]. Bringing together experts from fields such as human-computer interaction (HCI), design, algorithms, policy guidance, data protection law, and education, alongside AI practitioners, can harmonise diverse perspectives and terminologies to address the complexities of ethical AI for children. Collaboration should also be extended to other regulators at the international level such as the Global Online Safety Regulators Network and the Global Privacy Assembly, to establish best practices across jurisdictions and work towards creating global standards for online safety and privacy[301].

---

[296] Wang, G., Sun, K., Atabey, A., Pothong, K., Lin, G. C., Zhao, J., & Yip, J. (2023, April). Child-Centred AI Design: Definition, Operation, and Considerations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

[297] Wood, S. (2024). Impact of regulation on children's digital lives.

[298] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, *6*(3), 265-270.

[299] Wood, S. (2024). Impact of regulation on children's digital lives.

[300] Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, *6*(3), 265-270.

[301] Wood, S. (2024). Impact of regulation on children's digital lives.

Regulatory experimentation tools[302][303]**:** Regulatory experimentation tools, such as sandboxes, can be used to provide a controlled environment where innovative safety mechanisms, such as AI-based content moderation and child-specific safety features, can be tested and refined. By testing localised and adaptive GenAI models, regulators can gather valuable insights to inform future regulations, making them more responsive and tailored to evolving online safety challenges.

Quantifying harms: Regulators can develop robust methods for quantifying the potential harms children may encounter, with a focus on several key dimensions: the age group most affected, the severity of harm, the potential scale, and the current availability and effectiveness of harm reduction measures. **(1) Age groups:** different age groups experience varying degrees of vulnerability to online risks. Younger children, for example, may be more susceptible to exposure to inappropriate content, while older children may face heightened risks related to privacy violations or peer harassment; **(2) Severity of harm:** Severe harms, such as psychological trauma from cyberbullying or sustained exposure to harmful content, can have lasting effects on a child's development and well-being. The impact of harm must be measured not only by its immediate effects but also by its long-term consequences. **(3) Potential scale:** The potential scale of harm, or how widespread a particular risk could become, is another critical metric. Quantifying how many children are exposed to these risks, and under what conditions, helps gauge the urgency and scope of regulatory interventions. **(4) Harm reduction measures:** Existing harm reduction measures, such as content moderation tools, default privacy settings, and child-specific features, must also be evaluated for their effectiveness. The metrics can be developed through consultation workshops with child specialists, policy-makers, and domain designers using q-sorting methods for ranking harms[304].

5.3. For ecosystems (parents, teachers, caregivers, civil society, and public)

Everyday citizen audit/Decentralised content moderation: Users should be able to flag problematic content, with a special group of "trusted flaggers", such as individuals, tech-savvy NGOs, or volunteer coders. This approach leverages crowd-sourced efforts to monitor and correct GenAI output. The flaggers would register with the appropriate authority and act as a decentralised content monitoring team. They could test different prompts to detect harmful content, identify tools that bypass content moderation, and report to developers or deployers when issues arise[305].

Companies should have technical engineers teams working with developers or deployers, who are responsible for addressing flagged content. Notices from trusted flaggers should be prioritised by the content moderation team. These engineers would modify the AI system or block its output to prevent the flagged prompt from generating harmful content and identify ways to prevent potential misuse by malicious

[302] Eben, M., Erickson, K., Kretschmer, M., Cifrodelli, G., Li, Z., Luca, S., ... & Schlesinger, P. (2023). Priorities for generative AI regulation in the UK: CREATe response to the Digital Regulation Cooperation Forum (DRCF).
[303] Spatari, N. (2024, September). Exploring Regulatory Sandboxes: Safeguarding AI-Based Software for Minors. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 158-162).
[304] Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, *9*(1), 1-13.
[305] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).

actors. This blend of centralised and decentralised monitoring is expected to be more effective in managing harmful content such as hate speech and fake news generated by GenAI tools/systems[306].

Civil society watchdog: Under OSA, civil society cases can be filed under "super-complaints". Social media and search engine companies, including Meta, YouTube, and Snapchat, and Google have been criticised for slow and inadequate responses to victim-survivor requests regarding removing image-based sexual abuse (IBSA) and online harm contents[307]. Civil society organisations such as Revenge Porn Helpline has been acting as a bridge between victims and platforms and advocated for stronger and more efficient response mechanism from the platform companies. Regulators should continue collaborating with civil society organisations for evidence gathering and understanding of circumstances in which harms occurs.

Parent-AI collaborative system: The unpredictable and evolving nature of GenAI content requires a more dynamic and adaptive approach to content moderation and parental controls. Traditional systems based on fixed categories or age ratings are insufficient for managing the complexities of dynamically generated content. Instead, in the parent-AI collaborative system, parents can actively participate in evaluating, adjusting, and managing GenAI outputs[308]. This approach allows parents to continuously refine the content filtering process and ensure it meets their child's specific developmental needs and family values. Children can also be involved in the co-designing process, as it would reflect their perspectives and lead towards more effective and mutually agreed-upon solutions.

5.4. For children

Children can play different roles as **users, contributors and innovators** in shaping AI policies and practices. Mathiyazhagan & La Fors (2023)[309] presents different child participation models in the AI development process:

Table 1: Child participation models in the AI development process

Children role	Planning	Data collection	Data access	Use of algorithms	Deployment	Reporting and dissemination
User	No power	No other commerciall y viable option to use services but agreeing to	Free access for big tech and no power to user	No control, no knowledge and no power	No understanding of targeted audience and consequences	No control of representation

[306] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).
[307] Huber, A. R., & Ward, Z. (2024). Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act. *European Journal of Criminology*, 14773708241255821.
[308] Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2024). Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications. *arXiv preprint arXiv:2406.10461*.
[309] Mathiyazhagan, S., & La Fors, K. (2023). Children's right to participation in AI: Exploring transnational co-creative approaches to foster child-inclusive AI policy and practice. *Information Polity*, *28*(1), 141-153.

		and sharing data				
Contributor	Limited engagement of stakeholders and power	Tokenism/limited power to share opinions regarding data collection	Engagement is embraced to the extent to which it provides economic surplus	Children's coding is embraced to the extent to which it results in economic surplus	No power but more knowledge of the consequences of data processing and algorithms	Possible misrepresentation /tokenism
Innovator (designer, coder, UX developer)	Children gain power through meaningful engagement	Understanding and more assertiveness to shape how/who's data is collected	More assertiveness to shape how/who's data is accessed/under what conditions shared	More understanding of how the algorithm works	More power to control where and when to be applied	Offer a better representation of social diversity

AI literacy[310]: As shown by previous examples, educational activities play a vital role in demystifying AI systems and helping children develop a clear understanding of both their capabilities and limitations. Age-appropriate ethics education is critical to helping children potential harms, biases, and privacy risks in their interactions with GenAI. This education also fosters critical thinking, encouraging children to question AI-generated content, avoid over-reliance on AI and unhealthy attachment to GenAI.

Youth-shaped algorithms/Child-appropriate platform moderation[311]: Young people could be actively involved in algorithm design and audits, helping to report harmful materials while prioritising uplifting and positive content. A child-friendly feedback loop should be established where children can report harmful or uncomfortable experiences with AI directly to platform moderators or developers. Children can also exercise control over their digital experience by having the option to opt-in to data-driven algorithms, with clear choices about how their data is collected and used. A more inclusive and child-appropriate platform moderation is needed, sensitive to age and diversity.

VI. CONCLUSION

This report highlights the transformative potential and inherent risks of GenAI chatbots for children. While these technologies offer substantial benefits in terms of educational support, creativity enhancement, inclusivity for children with special needs, and online safety, they also pose significant risks, including

[310] Su, J., Ng, D. T. K., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence*, *4*, 100124.
[311] Reich, S. M., Starks, A., & Su, Z. (2024). One (Adult) Size Does Not Fit All: The Importance of Development in Digital Design and Utilization. *Youth Wellbeing in a Technology Rich World*.

exposure to harmful content, misinformation, emotional dependency, and privacy concerns. Given that children are particularly vulnerable to these risks due to their ongoing cognitive and social development, it is essential to adopt a cautious and comprehensive approach to deploying GenAI chatbots in environments accessible to young users.

The findings emphasises the need for a comprehensive and holistic regulatory framework that includes independent audits, transparency obligations, child-centric design principles, and close collaboration across AI, legal, and psychological domains. Furthermore, child-specific considerations in AI governance, like developmental stages and social contexts, must be integrated into policy-making to ensure that the technology supports safe, inclusive, and empowering interactions.

As the GenAI landscape continues to evolve, stakeholders, including developers, regulators, educators, parents, and civil society, must work together to ensure that GenAI tools are not only safe but also contribute positively to children’s learning and development. This report highlights the importance of ongoing research, interdisciplinary collaboration, and the active inclusion of children's perspectives in the design and regulation of GenAI technologies, aiming for a balanced approach that maximises benefits while minimising risks.