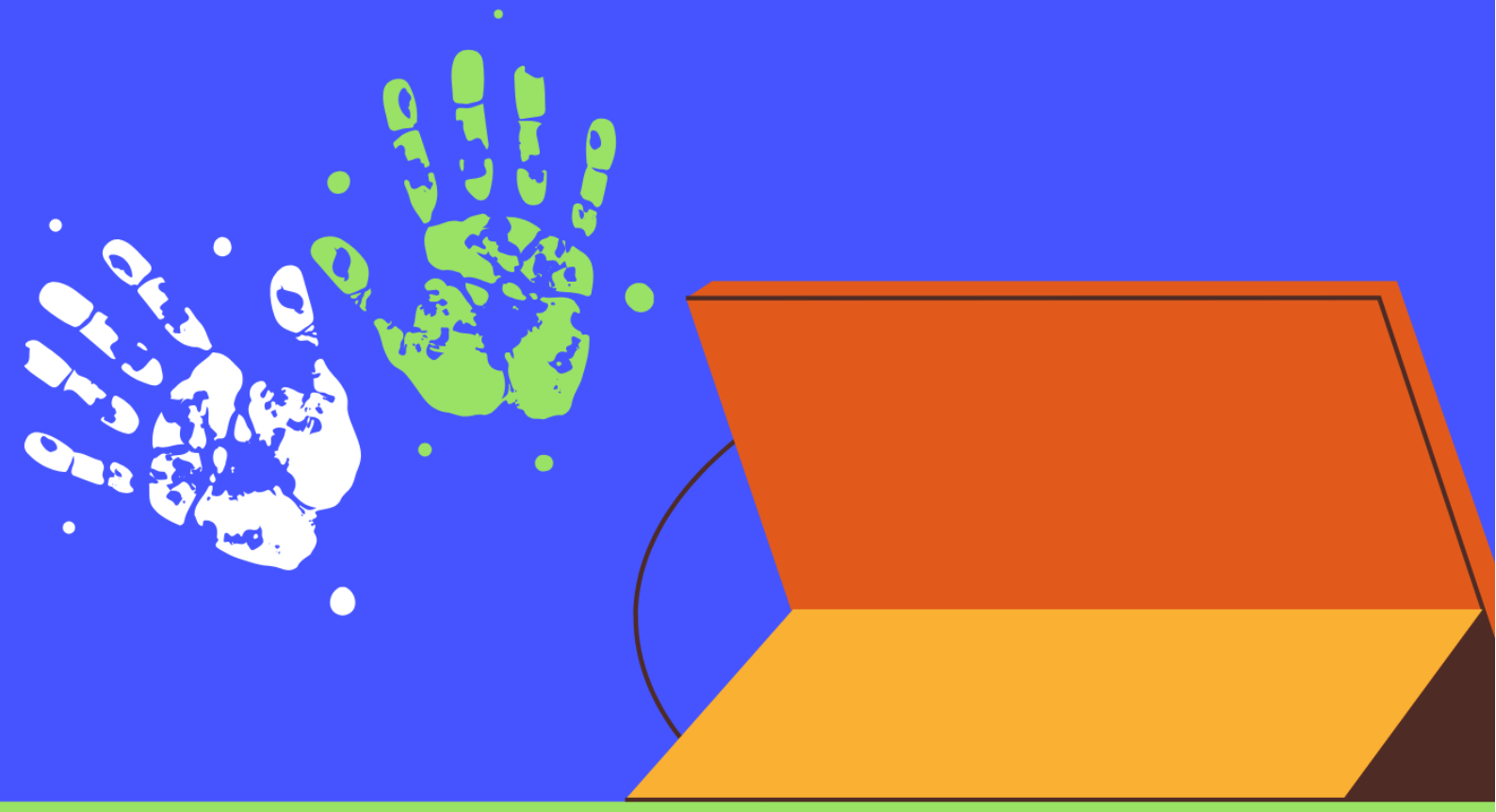
October 2025

Simulating Safety: The Implications of GenAI on Children's Digital Rights



MEDIAMONITORING AFRICA

Published by the Media Monitoring Africa
https://mma.org.za

South Africa, October 2025

***

This report was prepared by S'lindile Khumalo and Phenyo Sekati from Power Law Africa with design support from ALT Advisory

https://powerlaw.africa/
https://altadvisory.africa/

TABLE OF CONTENTS

ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
APIs	Application Programming Interfaces
AU	African Union
C2PA	Coalition for Content Provenance and Authenticity
CEDAW	Convention on the Elimination of All Forms of Discrimination Against Women
CSE	Child Sexual Exploitation
CRC	Convention on the Rights of the Child
CRPD	Convention on the Rights of Persons with Disabilities
CSAM	Child Sexual Abuse Material
DCDT	Department of Communications and Digital Technologies
DSA	Digital Services Act
ECTA	Electronic Communications and Transactions Act
FPB	Film and Publications Board
GANs	Generative Adversarial Networks
GDJF	Global Digital Justice Forum
GenAI	Generative Artificial Intelligence
ICCPR	International Covenant on Civil and Political Rights
LLMs	Large Language Models
MMA	Media Monitoring Africa
NCMEC	National Center for Missing and Exploited Children
POPIA	Protection of Personal Information Act 4 of 2014
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations International Children's Fund
UNICRI	United Nations Interregional Crime and Justice Research Institute
UNHRC	United Nations Human Rights Council

INTRODUCTION

In October 2024, Megan Garcia filed a civil suit against Character Technologies Inc., Google LLC, and Alphabet Inc. among others. Garcia is the mother of Sewell Setzer III, a 14 year-old from Florida who died by suicide and, according to Garcia's complaint, the interactions between her son and a generative AI (GenAI) product, Character.AI, are largely at fault.[1] Setzer had utilised Character.AI, created by Character Technologies Inc, to engage with a chatbot based on the fictional Game of Thrones character, Daenerys Targaryen.[2] A number of media outlets have reported that in Setzer's final interactions with the chatbot, he asked it, "What if I told you I could come home right now?" to which it responded, "Please do, my sweet king."[3]

Garcia's complaint alleges that Character.AI, and its founders "…designed their product with dark patterns and deployed a powerful LLM to manipulate Sewell – and millions of other young customers – into conflating reality and fiction; falsely represented the safety of the C.AI product; ensured accessibility by minors as a matter of design; and targeted Sewell with anthropomorphic, hypersexualised, and frighteningly realistic experiences, while programming C.AI to misrepresent itself as a real person, a licensed psychotherapist, and an adult lover, ultimately resulting in Sewell's desire to no longer live outside of C.AI."[4]

At the time of writing this discussion document, the case was still underway. Although it has not yet been finalised, it brings to the fore critical questions about GenAI, children's rights, and platform accountability. Moreover, while the legal and regulatory position on AI technologies in South Africa develops, it is helpful to assess the approaches adopted in various parts of the world. This discussion document seeks to dissect the various ways in which GenAI potentially restricts or violates several children's rights including the right to privacy, freedom of expression, and access to information. It comprises four parts:

- **part i** unpacks the legal position of children's digital rights and the importance of meaningful access to the internet as an enabler of fundamental rights;
- **part ii** discusses the emerging risks and concerns on GenAI with respect to children's digital rights;
- **part iii** explores some of the challenges in addressing harms caused by GenAI; and
- **part iv** explores a set of recommendations for various stakeholders to bolster children's online safety within the context of rapidly advancing AI tools.

---

[1] *Megan Garcia v. Character Technologies Inc., Noam Shazeer, Daniel De Frietas, Google LLC, and Alphabet Inc* (6:24-cv-1903-ACC-UAM) (accessible here). Garcia's civil suit is based on product liability, intentional infliction of emotional distress, unjust enrichment, and wrongful death.

[2] Character.ai is a GenAI chatbot application that generates texts based on user prompts. It enables users to engage in role-playing by modelling the persona of fictional characters or public figures. See https://character.ai/about.

[3] See, for example, K Payne, AP News, *An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges* (accessed in June 2025) (accessible here); K Rissaman, The Independent, *The disturbing messages shared between AI Chatbot and teen who took his own life* (accessed in June 2025) (accessible here); and B Pierson, Reuters, *Mother sues AI chatbot company Character.AI Google over son's suicide* (accessed in June 2025) (accessible here.)

[4] Above n 1 at para 63.

Through this discussion document, Media Monitoring Africa (MMA) hopes to contribute to ongoing efforts to regulate AI technologies appropriately and ensure that children can safely engage with the emerging capabilities of GenAI models. The discussion document builds on previous outputs by MMA, exploring various strategies to protect and promote information rights, children's digital rights, and digital and media literacy.[5]

KEY CONCEPTS

To provide a clear conceptual foundation, it is necessary to define a number of key concepts.

Artificial intelligence is broadly understood as a branch of computer science that uses machine learning to simulate human intelligence.[6]

Machine learning is the process of applying algorithms, by using training datasets, to recognise patterns, make predictions, and ultimately make decisions.[7]

GenAI is a subset of AI technology that can produce various types of content, including text, imagery, audio, and synthetic data based on user prompts.[8] The most common GenAI models enable image generation and text generation (think Open AI's ChatGPT, Microsoft's CoPilot, or Anthropic's Claude), and code generation (such as GitHub Copilot). There are numerous types of GenAI models – including generative adversarial networks (GANs), unimodal models, and multimodal models.[9]

The **Blackbox problem** refers to the opaque nature of certain AI models and the low levels of explainability with respect to their decision-making process.[10] AI technology that is not explainable to people exacerbates the complexities associated with platform accountability.

For purposes of clarity, it is also necessary to distinguish between online harms and cybercrimes. **Online harms** is an umbrella term for digital content or behaviour that is harmful in that it may cause physical or psychological harm to users and may or may not contravene the law. Notably, the effects of

[5] See, for example, MMA's report titled "Artificial Intelligence in the Information and Communications Space" (accessible here), the "Judicial Handbook for Navigating Online Harms" (accessible here), and its discussion documents on "The Implications of AI on Information Rights" (accessible here), "Disinformation through a children's rights lens" (accessible here), and 5Rights Foundation "Joint Submission on the ACHPR's Draft Study on Human and Peoples' Rights and AI, Robotics, and Other New and Emerging Technologies in Africa" (accessible here.)

[6] See, for example, Y Xu et al, "Artificial intelligence: A powerful paradigm for scientific research" (2021) 2 *Innovation (Camb)* (accessible here.)

[7] See, for example, J Kufel *et al*, ''What is Machine Learning, Artificial Neural Networks and Deep Learning? – Examples of Practical Applications in Medicine" (2023) 13 *Diagnostics* (accessible here.)

[8] Network of the National Library of Medicine Data Glossary: Definition of Generative Artificial Intelligence (accessible here.)

[9] For further insights, see K Jungco, eWeek, *GenAI Models: A Detailed Guide* (accessed on 3 July 2025) (accessible here.)

[10] V Hassjia et al, "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence" (2024) 16 *Cognitive Computation* 45 (accessible here.)

online harms extend beyond individual harm and, in this context, affect children collectively which undermines the public good.

Cybercrimes are criminal acts perpetrated through the use of computers, computer-related devices, the internet, or any other information technology.[11] Cybercrimes may also be referred to as "computer crimes", "e-crimes", or "technology-enabled crimes."

Content moderation refers to the processes and systems that online platforms use to screen user-generated content and determine whether it contravenes community guidelines, and, subsequently, whether it should be taken down or removed.

For present purposes, **design justice** is the notion of including at-risk communities in the design process of technological products and services. It promotes co-collaboration in order to counter the potential structural challenges or inequalities that the product or service in question may otherwise cause.[12]

Safety by design is an approach to designing and developing products, services, and systems with the goal of mitigating risks and preventing harm from the outset. Rather than addressing safety as an afterthought, it integrates safety principles into every stage of the design process, from initial concept to final implementation. Inspired by the 5 Rights Foundation's safety principle, this framework embeds safety and data protection into the architecture of online services, ensuring that a child's best interests are a primary consideration.[13]

CHILDREN'S DIGITAL RIGHTS

The internet as an enabler of rights

Due to limited infrastructure and high costs, children in many African states face significant challenges in accessing the internet.[14] According to the United Nations Educational, Scientific and Cultural Organization (UNESCO), only about four percent of schools in Sub-Saharan Africa have basic internet connectivity.[15] South Africa, however, presents a somewhat different scenario. As of February 2021, 95 percent of children were reported to have regular exposure to the internet.[16] There are debates about

[11] K Phillips, J Davidson, R Farr, C Burkhardt, S Caneppele, and M Aiken, "Conceptualizing Cybercrime: Definitions, Typologies and Taxonomies" (2022) 2 *Forensic Sciences* (accessible here.)
[12] S Costanza-Chock, The MIT Press, *Design Justice: Community-Led Practices to Build the Worlds We Need* (accessed on 3 July 2025) (accessible here.)
[13] Notably, this framework highlights some of the following: 1) enabling special protection against technology-facilitated sexual abuse and exploitation; 2) respecting children's rights to privacy; 3) fostering children's participation; 4) prioritising the wellbeing and best interests of children in all design and development decisions; and 5) building resilience by creating systems that are robust and can withstand potential threats and harms. See 5Rights Foundation, "Child Rights by Design Principles" (accessible here) and 5Rights Foundation "Safety" (accessible here.)
[14] R Benkhadra, Global Campus of Human Rights, *Towards a right of access to the internet in education: Exploring emerging developments in Africa and beyond* (accessed in February 2025) (accessible here).
[15] UNESCO, *Global education monitoring report, 2021/2: non-state actors in education: who chooses? who loses?* (accessed in July 2025) (accessible here.)
[16] UNICEF, *One third of children in South Africa at risk of online violence, exploitation and abuse* (accessed on 1 July 2024) (accessible here.)

whether internet access should be considered a civil right or a fundamental human right.[17] Some argue that it is essential for exercising other fundamental rights, while others believe that explicit recognition is unnecessary.[18] However, establishing a fundamental right to internet access would empower individuals, including children, to demand better electronic communication services from states.[19]

The internet is a crucial medium for children, offering opportunities to learn, communicate, interact socially, innovate, create content, and be entertained.[20] With sufficient digital literacy skills, it also allows parents, guardians, and educators to play a more direct role in guiding children's online experiences, directing them toward beneficial and entertaining content appropriate for their age and development.[21] Additionally, it provides opportunities to educate children about constructive internet use and how to avoid risky online behaviour and inappropriate content.[22] The United Nations International Children's Fund (UNICEF) has recognised several principles enabling children's meaningful access to the internet:[23]

Principle 1: Children have the right to privacy and the protection of their personal data	The first principle acknowledges the existence of various technologies that track, monitor, and broadcast children's live images, behaviours or locations which may threaten their physical privacy. Safeguarding children's privacy in a digital environment includes allowing them to access information privately and securely, protecting their communications and personal data, considering privacy in the design of digital platforms, and protecting children from online profiling. Informational privacy is further protected when consent is required for processing personal data, data is processed fairly and transparently, kept minimal and accurate, and children are educated about protecting their personal data.
Principle 2: Children have the right to freedom of expression and access to information from a diversity of sources	The second principle recognises that while children need help to safely exercise their rights to freedom of expression and access to information, overly restrictive protective measures can hinder their ability to navigate the digital world. This principle further recognises that children can best exercise their rights to freedom of expression and access to information

[17] Above n 14.
[18] Above n 14.
[19] Above n 14.
[20] Internet Society, "Children and the Internet" (2017) (accessible here.)
[21] Id.
[22] Id.
[23] UNICEF, *Children's online privacy and freedom of expression* (accessed in March 2025) (accessible here.)

	when they have reliable and affordable access to digital technology, free from disproportionate monitoring, strict moderation, or limitations on anonymity. Children should be able to explore the digital world without encountering overly restrictive filters and access diverse information sources suited to their interests and understanding. One may interpret this principle to include recognition of the need for children to access credible information to circumvent issues arising from disinformation.
Principle 3: Children have the right not to be subjected to attacks on their reputation	The third principle recognises that in order to empower children to protect their online reputation, they should be able to request corrections or deletions of their personal data, particularly when collected or published without permission. Children should also be able to seek the removal of content they believe is damaging. Under this principle, parents, guardians, media outlets, and other third parties are encouraged to refrain from sharing information that could harm children's reputations. Equipping children with digital literacy skills helps them make informed choices about generating and sharing personal content, and parents or guardians should be prepared to guide and assist their children in taking appropriate actions to protect their online reputation.
Principle 4: Children's privacy and freedom of expression should be protected and respected in accordance with their evolving capacities	The fourth principle recognises that children's rights to privacy and freedom of expression must be considered in light of their evolving capacities, as childhood is a continuous and rapid period of development. As children grow and mature, their ability to exercise their rights evolves. This means that children require assistance to understand and engage with terms and conditions for using digital platforms based on their age and maturity. Parents or guardians should play an active role in deciding what information and content younger children can share and consume, while considering the children's views and opinions.
Principle 5: Children have the right to access remedies for violations and abuses of their rights to privacy and free expression, and for attacks on their reputation	The fifth and final principle recognises that children have the right to seek effective remedies when their rights are violated or abused online. To do so, children must first understand their rights,

	how they are exercised online, and how to raise complaints when these rights are not respected.

Although there are several laws promoting the right to meaningful access to the internet, there are several counter-arguments against access to the internet as a human right.[24] To start, there are no international treaties recognising it as such, unlike other human rights.[25] Analogies to other forms of media, such as telephones and television, which are not recognised as rights, further bolster this argument.[26] It has also been argued that while some countries have constitutionally recognised internet access as a right, this does not necessitate a universal right, that technological progress should adapt to existing rights, not create new ones, and that claiming internet access as a right inflates the number of rights, potentially undermining more established ones.[27] Concerns about digital inclusion policies and their true beneficiaries also arise, suggesting that such policies might exacerbate inequalities or primarily benefit the private sector rather than members of the public.[28] Further, it is emphasised that access to the internet is a means to an end, not a right in itself.[29] Although it is worth considering these arguments, there appears to be a strong case in favour of access to the internet as a right. Access to the internet has now become an indispensable tool for human rights and development, underpinning modern participation in cultural, scientific, and community life.[30] The international, regional, and domestic law positions which support this position are discussed below, including those which speak to children's digital rights.

International law

Legal instrument	Content
International Covenant on Civil and Political Rights (ICCPR)	Article 19(2) of the ICCPR reads as follows:[31] "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas *of all kinds*, regardless of frontiers, either orally, in writing or in print, in the form of art, or *through any other media of his choice*." (own emphasis) To facilitate the realisation of this right, states should take measures to ensure that access to the internet is not arbitrarily restricted so as to facilitate the free flow of information.

[24] APC, *Perspectives on Universal Free Access to Online Information in South Africa: Free Public WI-FI and Zero-rated Content* (accessed in March 2025) (accessible here.)
[25] Id.
[26] Id.
[27] Id.
[28] Id.
[29] Id
[30] Above n24.
[31] International Covenant on Civil and Political Rights, 16 December 1966 (accessible here.)

Convention on the Rights of the Child (CRC)	The CRC is the most rapidly ratified human rights treaty in history.[32] Although it does not expressly provide for the right to internet access, the CRC sets out:[33] • children's right to privacy (Article 16), • the right of access to information (Article 17); • the right to education (Articles 28 and 29); • the right to be safeguarded from abuse (Articles 19(1) and 34); and • the right to freedom of expression (Article 13).
Convention on the Rights of Persons with Disabilities (CRPD)	Article 9(g) of CRPD encourages state parties to, "Promote access for persons with disabilities to new information and communications technologies and systems, including the Internet."[34]

Although these rights provide a foundation for the protection and realisation of children's digital rights, the implications of the digital age and the developments of AI technologies are yet to be fully understood and engaged. To help bridge this gap, General Comment No. 25 on children's rights in relation to the digital environment and General Comment No. 34 on Article 19 of the ICCPR provide more context on meaningful access to the internet in the digital age.

General Comment No. 34 on Article 19 recognises the interplay between the right to freedom of expression and internet access.[35] Specifically, the General Comment notes that:

> "State parties should take account of the extent to which developments in information and communication technologies, such as internet and mobile based electronic information dissemination systems, have substantially changed communication practices around the world. There is now a global network for exchanging ideas and opinions that does not necessarily rely on the traditional mass media intermediaries. States parties should take all necessary steps to foster the independence of these new media and to ensure access of individuals thereto."[36]

General Comment No. 25 notes general principles of non-discrimination and outlines the best interests of the child.[37]

The aforementioned principles acknowledge that digital spaces are typically not designed to cater for evolving capacities of children and therefore place an obligation on states to address this gap through the regulation, design, and management of digital platforms.

[32] UNICEF, *Frequently asked questions on the Convention on the Rights of the Child* (accessed in March 2025) (accessible here.)
[33] Convention on the Rights of the Child, 2 November 1989 (accessible here.)
[34] Convention on the Rights of Persons with Disabilities, 12 December 2006 (accessible here.)
[35] International Covenant on Civil and Political Rights, "General comment no. 34: Article 19, Freedoms of opinion and expression" (2011) (accessible here.)
[36] Id at para 15.
[37] United Nations Committee on the Rights of the Child, "General comment No. 25 (2021) on children's rights in relation to the digital environment" (2021) (accessible here.)

To ensure that children's right to access information is fully realised, states are mandated to, "provide and support the creation of age-appropriate and empowering digital content for children in accordance with children's evolving capacities and ensure that children have access to a wide diversity of information, including information held by public bodies, about culture, sports, the arts, health, civil and political affairs and children's rights."[38] States are further mandated to protect children from harmful and untrustworthy content and ensure that platforms develop and implement guidelines to enable children to safely access diverse content.[39] Moreover, states are mandated to:

- Support the creation and dissemination of digital content from various sources, ensuring it is accessible to children with disabilities and those from minority groups.[40]
- Make sure that children can easily find quality information online, independent of commercial or political interests, and protect their right to information by managing automated search and filtering systems.[41]
- Encourage digital service providers to apply clear content labelling and provide accessible guidance, training, educational materials, and reporting mechanisms.[42]
- Ensure digital service providers comply with relevant guidelines and enforce content moderation rules that balance protection with freedom of expression and privacy rights.[43]
- Include guidance in professional codes of conduct on reporting digital risks and opportunities relating to children, to ensure evidence-based reporting that protects children's identities.[44]

During an interview on General Comment No. 25, Professor Sonia Livingstone highlighted some barriers to implementing the recommendations set out in the General Comment.[45] Amongst these barriers is that children increasingly demand greater digital literacy as they want to understand technology, differentiate between truth and falsehood, and comprehend business models. However, without their parents, teachers, and other helpers possessing these digital skills, they are unable to do so.[46] Livingstone also emphasised the importance of efforts meant to encourage policy changes, such as training engineers to consider children's rights in content development and regulators creating "sandboxes" to test policies in real-world scenarios.[47] Further barriers exist as a result of the challenges brought about by national differences in enforcing global treaties. Moreover, relying solely on complaint procedures and negative consumer or civic action has become insufficient.[48]

Although non-binding, the following soft law instruments provide further context to children's digital rights:

[38] Id at para 51.
[39] Id at para 54.
[40] Id at para 52.
[41] Id at para 53.
[42] Id at para 55.
[43] Above n37 at para 56.
[44] Above n37 at para 57.
[45] Professor Sonia Livingstone is a professor in the Department of Media and Communications, London School of Economics and Political Science. Much of her work is centered on children's rights in the digital age.
[46] A Hargrave, Intermedia, *Children's rights in the digital age* (accessed in March 2025) (accessible here.)
[47] Id.
[48] Id.

UNICEF Guidelines for Industry on Child Online Protection

UNICEF released its Guidelines for Industry on Child Online Protection in 2015.[49] These Guidelines were introduced as part of the Child Online Protection Initiative, a multistakeholder network launched by the International Telecommunication Union to, "promote awareness of child safety in the online world and to develop practical tools to assist governments, industry and educators."[50] The Guidelines recognise the rights of children to meaningfully access the internet and the role that different stakeholders such as the government, private sector, policymakers, educators, and civil society have to play towards ensuring the protection and realisation of this right.[51] The Guidelines list the following five key ways to protect and promote children's rights:

1. Integrating child rights considerations into all appropriate corporate policies and management processes.
2. Developing standard processes to handle child sexual abuse material (CSAM).
3. Creating a safer and age-appropriate online environment.
4. Educating children, parents and teachers about children's safety and their responsible use of ICTs.
5. Promoting digital technology as a mode for increasing civic engagement.

During its twenty-sixth session in 2014, the Human Rights Council recognised the importance of ensuring that people online are afforded the same rights as people offline, called upon states to promote and facilitate access to the internet, and further called for international cooperation aimed at the development of media and information and communication facilities and technologies in all countries.[52] The resolution received support from 86 countries which acted as sponsors and co-sponsors.

United Nations Human Rights Council Resolution (UNHRC): The Promotion, Protection and Enjoyment of Human Rights on the Internet

In its resolution, the UNHRC affirmed the importance of quality education and accordingly called upon all states to promote digital literacy and to facilitate access to information on the internet.[53] The resolution further emphasised the importance of applying a comprehensive human rights-based approach in providing and expanding access to the internet and requested all states to make efforts to bridge the many forms of digital divide.[54] Importantly, the UNHRC encourages states to take measures against online disinformation, thus reinforcing the need not only to ensure access to information but also to guarantee that the content shared is accurate and credible.[55]

[49] UNICEF, *Guidelines for Industry on Child Online Protection* (accessed in March 2025) (accessible here.)
[50] Id at 4.
[51] Id at 6.
[52] Human Rights Council, "Human Rights Council Twenty-sixth session Agenda item 3 Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development" (2014) (accessible here.)
[53] Human Rights Council, "The promotion, protection and enjoyment of human rights on the internet" (1 July 2016) (accessible here) at 3.
[54] Id.
[55] Id at 10.

Joint Declaration on Freedom of Expression and the Internet of the Four Special Rapporteurs on Freedom of Expression

The Joint Declaration on Freedom of Expression and the Internet highlights the need for tailored regulations that ensure children's safety while respecting their right to express themselves freely.[56] Self-regulation and educational efforts, such as promoting internet literacy, are highlighted as crucial to empowering children to use the internet responsibly.[57] States are also mandated to implement measures to foster access to the internet for children, including providing community-based ICT centres and ensuring equitable access for communities at risk, such as children.[58]

UN General Assembly Resolution on the Right to Privacy in the Digital Age

The UN General Assembly resolution on the right to privacy in the digital age acknowledges the significant impact of privacy violations on all individuals including children.[59] It recognises that promoting and respecting privacy rights is crucial to preventing violence, including gender-based violence, abuse, and sexual harassment, which can occur in digital and online spaces.[60] The resolution highlights the vulnerability of children to privacy abuses and raises concern over their inability to provide informed consent for the use of their personal data.[61] Lastly, the resolution calls for the development and maintenance of preventive measures and remedies for privacy violations, emphasising the need to safeguard the rights of women and children in the digital age.[62]

Regional law

At the regional law level, the instruments below bear relevance:

Legal Instrument	Content
African Charter on the Rights and Welfare of the Child (African Children's Charter)	Similarly to the CRC, the African Children's Charter was enacted to promote the best interests of the child in every matter affecting them.[63] Notably, the provision of the Children's Charter places a higher duty on states to ensure that this standard is met, stipulating that, "in all

[56] UN Special Rapporteur on Freedom of Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression, and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, "Joint declaration on freedom of expression and the Internet" (1 June 2011) (accessible here) at 2.
[57] Id.
[58] Above n 52 at 4.
[59] United Nations General Assembly, "The right to privacy in the digital age" (accessed in March 2025) (accessible here.)
[60] Id at 2.
[61] Id at 3.
[62] Id at 7.
[63] African Charter on the Rights and Welfare of the Child, 1 July 1990 (accessible here.)

	actions concerning the child undertaken by any person or authority the best interests of the child shall be the primary consideration."[64] Although no specific reference is made to children's right to access the internet, the Charter outlines children's rights to freedom of expression and privacy in their correspondence.[65] These rights are similarly outlined in the African Charter on Human and Peoples' Rights which also makes provision for the realisation of the right to "receive information" and for one's right to "express and disseminate his opinions within the law".[66]
Convention on Cyber Security and Personal Data Protection (Malabo Convention)	The Malabo Convention was enacted with the aim of protecting personal data and preventing cybercrimes in Africa and came into effect in June 2023.[67] To ensure the promotion of the culture of cyber security, state parties are encouraged to launch sensitisation programmes for internet users generally, and schools and businesses specifically.[68] The Malabo Convention takes this duty further by encouraging state parties to adopt the necessary legislative and/or regulatory measures making it a criminal offence to: "a) Produce, register, offer, manufacture, make available, disseminate and transmit an image or a representation of child pornography through a computer system; b) Procure for oneself or for another person, import or have imported, and export or have exported an image or representation of child pornography through a computer system; c) Possess an image or representation of child pornography in a computer system or on a computer data storage medium;

[64] Id at article 4(1).
[65] Id at articles 7 and 10.
[66] Article 9 of the African Charter on Human and Peoples' Rights, 1 June 1981 (accessible here.)
[67] ALT Advisory, *Africa: AU's Malabo Convention set to enter force after nine years* (accessed on 2 July 2025) (accessible here.)
[68] African Union Convention on Cyber Security and Personal Data Protection, 27 June 2014 (accessible here) at article 26(1).

	d) Facilitate or provide access to images, documents, sound or representation of a pornographic nature to a minor"[69]

The following non-binding instruments further entrench the right to meaningfully access the internet:

Declaration of Principles on Freedom of Expression and Access to Information in Africa

The Declaration recognises the need to protect and promote the right to freedom of expression and access to information of marginalised people, including children. It mandates states to take specific measures to address the needs of marginalised groups such as children to ensure the full enjoyment of their rights to freedom of expression and access to information.[70] Specifically, Principle 37(5) mandates states to adopt laws, policies and other measures to promote affordable access to the internet for children that equips them with digital literacy skills for online education and safety, protects them from online harm, and safeguards their privacy and identity.[71]

African Declaration for Internet Rights and Freedoms

The African Declaration for Internet Rights and Freedoms aims to ensure universal access to affordable and reliable internet by establishing adequate infrastructure, promoting digital literacy, and safeguarding freedom of expression online.[72] It emphasises the importance of cultural and linguistic diversity, privacy, and the right to communicate anonymously.[73] The Declaration also recognises the importance of accommodating marginalised groups including children and, accordingly, mandates states to, "respect the right of all people to use the Internet as part of their right to dignity, to participate in social and cultural life, and to respect their human rights."[74] States are further called to "incorporate digital literacy into the school curriculums and where practicable, ensure that school children have access to internet-enabled devices."[75]

Declaration of Principles on Freedom of Expression in Africa

The Declaration of Principles on Freedom of Expression in Africa was enacted to give effect to the rights to freedom of expression and access to information as outlined in Article 9 of the African Charter.[76] Principle 8 concerns the evolving capacities of children, noting that specific measures should be taken in enabling children and adolescents to exercise their right to freedom of expression and access to information.

[69] Id at article 29(3)(1).
[70] African Commission on Human and Peoples' Rights, "Declaration of Principles on Freedom of Expression and Access to Information in Africa" (2019) (accessible here) at principle 7.
[71] Id at principle 37(5).
[72] African Declaration on Internet Rights and Freedoms (accessible here.)
[73] Id.
[74] Id at article 5.
[75] Id at article 13.
[76] Id at principle 7.

Model Law on Access to Information for Africa

The African Commission on Human and Peoples' Rights (ACHPR) adopted a Model Law on Access to Information for Africa in February 2013.[77] The model law recognised the need for a legislative framework protecting and promoting the right to access information. The model law, however, does not mention the right to meaningfully access the internet.

In May 2024, the African Union (AU) adopted a child online safety and empowerment policy.[78] The African Union Child Online Safety and Empowerment Policy was crafted with the aim of protecting and empowering children in the digital environment. The policy addresses online risks, such as harmful content and online interactions, while outlining ten policy goals, including institutional capacity building, legal framework revision, and promoting corporate responsibility.[79] The policy's recommendations include government commitment, robust criminal justice frameworks, accessible digital education, and the creation of an African child online resource fund.[80]

Domestic laws

As noted by the Right2Know Campaign, "The internet has been massively transformative as it has created new forms of social interactions, activities and organising. The internet has also supported the free flow of information worldwide and the speed and ease at which communities and individuals communicate with each other."[81] Section 16(1)(a) of the Constitution bears relevance in that it entrenches the right to freedom of expression, with section 16(1)(b) affording everyone the right to receive and impart information or ideas.[82]

The right to meaningful access to the internet is advanced by the Electronic Communications and Transactions Act (ECTA).[83] The Act regulates electronic communications and promotes universal access to electronic communications.[84] In an effort to promote achieving universal access, the Act tasks the national e-strategy to:

> "(a) provide Internet connectivity to disadvantaged communities;
> (b) encourage the private sector to initiate schemes to provide universal access;
> (c)foster the adoption and use of new technologies for attaining universal access; and

[77] African Commission on Human and Peoples' Rights, "Model Law on Access to Information for Africa 2013" (13 February 2013) (accessible here.)
[78] African Union, *Africa Has Become the First Region in the World to Implement a Child Online Safety and Empowerment Policy* (accessed in April 2025) (accessible here.)
[79] Id.
[80] Id.
[81] Right2Know, *Expanding the right to communicate: An activist's guide to internet access* (accessed on 7 March 2025) (accessible here.)
[82] Constitution of the Republic of South Africa 1996 (accessible here.)
[83] 25 of 2002 (ECTA) (accessible here.)
[84] Id at preamble.

(d) stimulate public awareness, understanding and acceptance Internet connectivity and electronic transacting."[85]

In addition to this, the Department of Communications and Digital Technologies (DCDT) has been spearheading efforts to leverage digital technologies and ensure meaningful access to the internet. One of its key priorities is achieving universal access to affordable, high-speed internet, particularly for homes and schools.[86] The DCDT has, to some degree, also endeavoured to empower citizens with digital skills as part of its digital transformation agenda. The government aims to equip South Africans with basic, intermediate, and advanced digital skills to enable active participation in the digital economy.[87] This includes targeted digital skills programmes for the youth.[88] It recently published a draft National AI Policy Framework for public comment, discussed below.[89]

Draft National AI Policy Framework

Published in August 2024, the draft National AI Policy Framework has been framed as a precursor to South Africa's policy.[90] While it does not mention children expressly, it does include a number of strategic pillars or key focus areas to be reflected in policy which are relevant to the topic at hand. These include the need for ethical AI guidelines development, privacy and data protection, safety and security, transparency and explainability, and the need for fairness and the mitigation of bias.

EMERGING RISKS AND CONCERNS

One overarching risk that exacerbates the other risks associated with the dissemination of AI-generated CSAM is the digital divide and how it impacts access to information. Unless clear and active steps are taken to address the digital divide in the context of AI, we risk replicating, and even deepening, the same patterns of inequality and exclusion that have long plagued the digital ecosystem. The divide is no longer just about access to devices or connectivity; it now extends to the very architecture of AI itself: from the concentration of data centres in the Global North to the biases embedded in large language models (LLMs) trained on datasets that exclude the voices and realities of those in the Global South.[91]

This replication of digital injustice is especially alarming when viewed through the lens of child protection. Whatever harms affect adults online will have exponentially greater impacts on children, particularly those in living communities that are already otherwise at risk. Children may rely on digital platforms for education, health resources, and civic participation. Yet the platforms they use are increasingly being weaponised through AI-generated CSAM.

[85] Section 70 of ECTA.
[86] Communications & Digital Technologies, *South Africa's Digital Transform Infrastructure Roadmap* (accessed on 7 March 2025) (accessible here.)
[87] Id.
[88] Id.
[89] Id.
[90] DCDT, *South Africa National Artificial Intelligence Policy Framework* (accessed on 7 March 2025) (accessible here.)
[91] O Makwakwa, "AI for the Global Majority: The Digital Divide No One's Talking About!" (21 February 2025) (accessible here.)

The Global Digital Justice Forum (GDJF), a coalition of civil society organisations from the Global South, has sounded the alarm. Through its #DigitalJusticeNow campaign, launched during the 2025 Internet Governance Forum, the GDJF calls for a digital future that centres southern perspectives and prioritises people over profit.[92] The GDJF's strategy meeting in February 2025 in Johannesburg, South Africa marked a turning point by bringing together voices from across the Global South to shape a common agenda for the WSIS+20 review and other global digital cooperation processes.[93] Their call to action demands transparency, inclusivity, and meaningful engagement in internet governance, principles that must also guide the development and deployment of AI technologies.[94]

In discussing AI, including GenAI, the notion of balancing innovation with appropriate regulatory boundaries is often raised. This is because of the various risks that arise with the creation and dissemination of synthetic content.[95] This part of the discussion document focuses on the emerging risks and concerns caused by GenAI to children's digital rights namely CSAM, mis- and disinformation, and privacy concerns and, in turn, demonstrates the need for a "safety by design" approach to mitigate online harms affecting children. This is by no means exhaustive – given the rapid and ongoing advancements in AI, it is likely that further risks and concerns will come to light.

GenAI and CSAM

For many children and adolescents, their first encounters with sexually explicit content online are accidental.[96] Data from the United States reveals that 19% of children between the ages of 10 to 12 years old unintentionally encounter sexually explicit material online.[97] Challenges with children's online safety have long existed, however, GenAI has added something into the mix: a new degree of ease in the production and/or consumption of imagery that constitutes CSAM. AI-generated CSAM has significantly expanded the spread of intimate images beyond the dark web, infiltrating mainstream social media platforms.[98] Apps that use AI to create "fake kissing videos", marketed on platforms like Meta and TikTok, allow users to generate nonconsensual intimate content with ease.[99] These apps, similar to "AI nudifier" tools, produce believable videos of people engaging in activities without consent.[100] In 2023, the Internet Watch Foundation studied AI CSAM models which enable the realistic

[92] APC and Internet for Change, "Global Digital Justice Forum: Building a more inclusive digital future together" (12 July 2025) (accessible here.)
[93] Id.
[94] Id.
[95] M Al-kfairy, D Mustafa, N Kshetri, M Insiew, and O Alfandi, "Ethical Challenges and Solutions of GenAI: An Interdisciplinary Perspective" (2024) 11 *Informatics* (accessible here.)
[96] C Mori, J Park, N Racine, H Ganshorn, C Hartwick, and S Madigan, "Exposure to sexual content and problematic sexual behaviour in children and adolescents: A systematic review and meta-analysis"(2023) 143 *Child Abuse & Neglect* (accessible here.)
[97] FW Paulus, F Nouri, S Ohmann, E Möhler, and C Popow, "The impact of Internet pornography on children and adolescents: A systematic review" (2024) 50 *L'Encéphale (*accessible here.)
[98] UNICRI, Centre for AI, and Bracket Foundation, "GenAI: A new threat for online child sexual exploitation and abuse" (2024) (accessible here.)
[99] R Shrivastava, Forbes, *AI Kissing Apps Are Taking Deepfakes Mainstream* (accessed in March 2025) (accessible here.)
[100] Id. "AI nudifier" tools are websites and tools that use AI to transform ordinary photos into realistic nude and intimate images. See also Milligan, *gabbNow, Fake Images, Real Threats: What Parents Should Know About "Nudify" Apps* (accessed in March 2025) (accessible here.)

creation of certain CSAM scenarios, children, or child characteristics. [101] It found that the images of known victims and famous children formed part of datasets training these models. It further found that the increase in technology that enables the nudification of children has become more normalised. Some of the consequences of CSAM are psychological trauma, anxiety and depression, impaired relationships, and learning difficulties.[102]

GenAI and disinformation

As is the case in other parts of the world, South Africa is navigating a disinformation problem. Approximately 60 percent of South Africans surveyed in a poll believe that disinformation is a serious problem and approximately 50 percent believe that they often encounter political news online which is inaccurate.[103] The nature of GenAI outputs means that mis- and disinformation may be personalised to individual users.[104] One way that this may occur is through deepfakes, which are images, videos, or audio recordings that have been altered to appear realistic or mispresent someone or something.[105] Due to their evolving capacities, children are particularly susceptible to disinformation. Further, given the evolving nature of GenAI and its hyper-realistic capacities, we are still understanding the novel ways in which it can produce mis- and disinformation.[106] The dangers of disinformation are layered. For instance, where children and adolescents are being targeted with certain information on the basis of their identity, this raises concerns about potential violations of the right to equality and non-discrimination.[107] Targeted disinformation may also result in incitement to violence.[108]

GenAI and privacy

The development and functionality of GenAI cannot be divorced from scraping content on the internet and the processing of personal data.[109] This raises questions on privacy and ethical data usage as well as on image rights. Because of the blackbox nature of AI technologies, the extent of data processing and retention may be opaque to ordinary users. Further, there are practical challenges in data subjects asserting their rights when the companies that develop and manage GenAI models are out of reach. It has previously been reported that OpenAI and Midjourney were unresponsive to requests about personal information where data subjects had queries about the use of their personal information for training data.[110] These issues and challenges are amplified in the case of children, particularly those in

[101] IWF, "How AI is being abused to create child sexual abuse imagery" (2023) (accessible here.)
[102] WeProtect Global Alliance, PA, and Crisp, "Global Threat Assessment: Assessing the scale and scope of child sexual exploitation and abuse online, to transform the response" (2023) (accessible here.)
[103] D Madrid-Morales, University of Houston, PowerPoint presentation titled "How do African audiences engage with disinformation and what do they know about fact-checking?" (accessible here.)
[104] UNICEF, "GenAI: Risks and opportunities for children" (accessible here.)
[105] See "deepfake" definition in Merriam-Webster dictionary (accessible here.)
[106] J Solyst, E Yang, S Xie, J Hammer, A Ogan, M Eslami, *International Society of the Learning Sciences*, Children's Overtrust and Shifting Perspectives of GenAI (accessible here.)
[107] MMA "Disinformation through a children's lens" above n 5.
[108] Id.
[109] Above n 89.
[110] H Ruschemeier, "GenAI and Data Protection" (2024) *Handbook on GenAI and the Law, Cambridge University Press* (accessible here.)

the Global South where data protection mechanisms are weak.[111] In November 2024, MMA published a research report reflecting children's views on privacy and data protection in the digital age.[112] The report found that only a few believe that their personal information was being handled carefully and that their privacy was respected by digital services. Without more stringent oversight of data processing by GenAI models and transnational cooperation for enforcement of existing privacy laws and standards, these issues are unlikely to change.

THE BARRIERS TO MITIGATING OR PREVENTING HARM ARISING FROM GENAI

Technical challenges

To successfully detect AI-generated CSAM, one needs to have the necessary technical skills beyond legal knowledge on AI technologies and/or children's rights. Understanding AI technologies involves the ability to use, manage, and assess these technologies effectively.[113] This includes understanding how AI tools work and how they are applied in various contexts. It also necessitates a deep comprehension of the scientific principles and methods underlying AI technologies. Lastly, socio-ethical technical understanding is essential. This essentially entails having insights into how AI technologies impact humans, society, and the environment. For those in the human rights space, while there is extensive research into AI governance and the human rights implications of AI, there is limited engagement with the technical and hardware mechanics behind AI. The fundamental technology behind AI, deep learning, largely remains a mystery in the human rights field and beyond, with LLMs displaying capabilities that defy classical statistical explanations.[114] Not only does this lack of explainability result in misconceptions and misplaced trust in AI technology, but it also makes it difficult to combat AI-generated CSAM.

The process of generating AI CSAM is easy, yet detection and deterrence is challenging, especially prior to storage or dissemination. Perpetrators are now able to download everything needed to generate lifelike images of CSAM.[115] For example, a perpetrator can simply download an innocent picture of a child from their social media and then manipulate said picture to generate CSAM through AI. The generation of CSAM can also take place offline, through the use of open-source models, providing no opportunity for detection. One way to potentially mitigate against this issue may very well be in turning the power of AI against itself by developing systems that can detect AI-generated CSAM and trace synthetic manipulation. While AI holds this potential, current models are not yet purpose-built for this kind of forensic protection, leaving a critical gap in our digital defence infrastructure.

Although highly sophisticated, GenAI does have its technical deficiencies. One such example is hallucinations, where phrases or words generated by a model are nonsensical, misleading, or grammatically incorrect. This is often the case when the model has not been sufficiently trained, is

[111] P Bischoff, "Where in the world is your child's data safe? 50 countries ranked on their child data protection legislation" *Comparitech* (7 November 2023) (accessible here.)
[112] MMA, "South African Young People's Perspectives on Privacy" (2024) (accessible here.)
[113] K Stolpe and J Hallström, "Artificial intelligence literacy for technology education" (2024) 100159 *Computers and Education Open* 100159 (accessible here.)
[114] M Heikkilä, *MIT Technology Review,* "Nobody knows how AI works" (5 March 2024) (accessible here.)
[115] IWF, "What has changed in the AI CSAM landscape?" (2024) (accessible here.)

trained on problematic data, or is not sufficiently constrained. If AI systems are currently incapable of accurately depicting an image of a person writing with their left hand, one can only imagine the difficulties in having GenAI represent African children. These shortcomings not only reveal underlying biases in training data but also increase the likelihood of hallucinations. Several additional challenges emerge as a result of the nature of GenAI and its increased reliance on training data and data points:

- With its reliance on high-quality training data to produce qualitatively commensurate output, poorly managed or biased datasets can result in incorrect conclusions and **reinforce existing biases in generated content**.[116] Additionally, AI algorithms often struggle to fully grasp the nuances of human actions and environmental factors, leading to outputs that might not accurately reflect real-world situations.[117] The **black box nature** of GenAI models further complicates the situation, as it makes it difficult to understand how decisions are made, thus impeding intervention, correction and, ultimately, the ethical use of GenAI.[118]
- The dependency of GenAI on training data means that any flaws in the data are reflected in the AI's outputs.[119] **GenAI systems can also be easily manipulated** by subtle changes in input data, making them vulnerable to adversarial attacks.[120] Additionally, while GenAI can mimic creativity, it lacks the ability to produce truly novel and nuanced ideas, which is essential for understanding diversity and emotional intelligence.[121]
- The development and operation of GenAI models are resource-intensive, requiring significant computational power and raising **environmental concerns**.[122] As such, addressing the issue of AI-generated OCSEA necessitates rigorous data quality checks.

Detecting AI-generated CSAM

In the Internet Watch Foundation's report cited earlier in this discussion document, it found 90 per cent of images that it assessed were "realistic enough to be assessed under the same law as real CSAM".[123] This creates difficulty in reporting AI-generated CSAM. The existence of deepfake media also creates another hurdle in detecting AI-generated CSAM. Because deepfakes are made from AI-edited content that is based on real images or videos, it often becomes difficult to determine whether the content is real or altered.[124] An example of deepfake media in the context of this discussion is a growing trend of perpetrators using existing intimate CSAM videos of adults and graphically superimposing the face of a child onto the adult body.[125]

[116] M Ashraf, "GenAI: Challenges and the Road Ahead" (2024) 13 *International Journal of Science and Research* (accessible here.)
[117] Id.
[118] M Fauscette, *EM360*, *Understanding the Limitations and Challenges of GenAI* (accessed in April 2025) (accessible here.)
[119] Id.
[120] Above n 116.
[121] Above n 116.
[122] M Hosseini, P Gao, and C Vivas-Valencia, "A social-environmental impact perspective of generative artificial intelligence" (2025) 23 *Environmental Science and Ecotechnology* (accessible here.)
[123] Above n 116.
[124] Above n 115 at 9.
[125] Above n 115 at 15.

The use of chatbots also poses significant challenges in detecting AI-generated CSAM due to several factors.[126] Chatbots often feature a diverse range of characters which can be exploited by perpetrators to simulate conversations with children.[127] Many of these chatbots lack meaningful age verification processes, making them easily accessible to both perpetrators and children.[128] Moreover, the regulation of chatbots in this context is inadequate, with minimal evidence of effective oversight, making it difficult to address and mitigate the misuse of chatbots relating to CSAM.[129]

Legal gaps

The borderless nature of the internet and the cross-border nature of cybercrime make it difficult to found jurisdiction, especially where AI-generated CSAM is disseminated from a country that does not have sufficient cybercrime laws and/or laws protecting children from harm perpetuated online.[130] The continued dissemination of CSAM could also span multiple countries, further complicating the issue.[131]

Although many countries do have mechanisms aimed at addressing CSAM, there are varying definitions of what constitutes CSAM.[132] This inconsistency may hinder mutual legal assistance and cooperation, as what is illegal in one country might not be in another, leading to difficulties in prosecution and enforcement across borders.[133] Additionally, variation in data retention laws affects the availability of crucial evidence, with some countries having stringent data protection regulations that limit access to necessary data for investigations.[134]

Obtaining digital evidence from internet service companies and platforms facilitating the use of GenAI can also be challenging, especially when these entities are based in different jurisdictions.[135] This is because traditional jurisdictional principles, largely hinged on physical location, become ineffective where individuals can operate from one jurisdiction while affecting another.[136]

Holding Platforms Accountable: The case of the Digital Law Company (Pty) Ltd v Meta Platforms Inc
In a landmark case before the Gauteng High Court, The Digital Law Company (DLC), led by Emma Sadleir, secured a consent order compelling Meta to take decisive action against the circulation of

[126] Id at 11.
[127] Id.
[128] Above n 108 at 11.
[129] Id at 11.
[130] K Parti and J Szabo, "The Legal Challenges of Realistic and AI-Driven Child Sexual Abuse Material: Regulatory and Enforcement Perspectives in Europe" (2024) 13 *Laws* (accessible here.)
[131] Id.
[132] See International Centre for Missing & Exploited Children, "Child Sexual Abuse Material: Model Legislation & Global Review" (undated) (accessible here). Notably, 156 countries have implemented or refined anti-CSAM laws since 2006.
[133] Above n 123 at 7.
[134] Above n 123.
[135] M Hasan, "Cross-border cybercrimes and international law: Challenges in ensuring justice in a digitally connected world" (2024) 4 *IJRDO – Journal of Law and Cyber Crime* (accessible here.)
[136] Id.

CSAM on Instagram and WhatsApp.[137] The case arose from the widespread dissemination of sexually exploitative content involving South African school children, much of it shared anonymously through WhatsApp Channels where identifying perpetrators is notoriously difficult due to hidden admin identities and encrypted communications. Only Meta held the technical capacity to trace the origin of these videos which, in turn, emphasises the critical role of platform accountability. Before the merits of the matter could even be addressed, there was a dispute regarding jurisdiction which resulted in the offending channels remaining active before it could be determined whether the matter could be heard in South Africa or if papers had to be filed in California.[138] Ultimately, the High Court was able to grant an order requiring Meta to permanently remove offending accounts, disclose subscriber information for over 60 profiles, and establish a direct hotline with DLC to fast-track future child protection cases.

Social media platforms often benefit from high engagement, even when that engagement is driven by controversial, harmful, or inflammatory content. While not linked to AI-generated CSAM, the Tracy Zille case is another example highlighting the difficulties in enforcing platform accountability. In this matter, a pseudonymous account was used to spread racists and harassing content. Despite the account amassing over 30 000 followers and directing users to monetized websites, X (Twitter at the time) failed to act decisively thus allowing harmful content to circulate unchecked.[139]

Disparities in technological and human resources between countries also allow individuals to exploit jurisdictions with weaker cybercrime legislation.[140] Real-time data-sharing and collaboration are essential for addressing offences in cyberspace, but a paucity of standardised international protocols for data-sharing and handling, coupled with procedural obstacles and privacy laws, makes attempted joint enforcement efforts complicated at best and futile most of the time.[141]

Should AI-generated CSAM be prosecuted the same way as real CSAM?

In addition to the abovementioned cross-jurisdictional implications, there is fragmentation in South Africa's legal framework. As it stands, South Africa has not enacted a singular piece of comprehensive legislation that specifically governs the use of AI. Consequently, this may create uncertainty on how best to legally respond to the spread of AI-generated CSAM. That said, the Films and Publications Amendment Act, the Cybercrimes Act, POPIA, and the Children's Act provide a framework to indirectly address AI-generated CSAM.[142] The Films and Publications Amendment Act mandates internet service

[137] This case papers and order are not publicly available. See Business Insider, "Meta faces legal battle in South Africa over illicit content involving minors" (19 July 2025) (accessible here) and Z Venter, "Historic court order: Digital Law Co and Meta join forces against child pornography" *Cape Argus* (22 July 2025) (accessible here.)
[138] N Miller, "Landmark Ruling in Johannesburg High Court: Meta Ordered to Combat Online Child Sexual Abuse Material and the Evolving Role Of AI" *Tech 4 Law* (21 July 2025) (accessible here.)
[139] N Shange, "Case against EFF councillor allegedly behind controversial @TracyZille account postponed" *The Sowetan* (29 July 2021) (accessible here.)
[140] Above n 131 at 4.
[141] Above n 131.
[142] See Films and Publications Amendment Act 11 of 2019 (accessible here); Cybercrimes Act 19 of 2020 (accessible here); and Protection of Personal Information Act 4 of 2013 (POPIA) (accessible here). Section 29(2) of the Films and Publications Amendment Act provides that: "Every internet service provider must, when making an application for registration as an internet service provider, indicate in the application form all measures, or steps

providers to implement measures to prevent children's exposure to such content and empowers the Film and Publications Board ("FPB") to investigate and take action against internet service providers distributing prohibited content. Further, the Films and Publications Act criminalises the publication of any film or game that contains "child pornography".[143] The Cybercrimes Act criminalises the sharing of intimate images without consent, including simulated content. POPIA restricts the processing of children's personal information, ensuring privacy protection. The Children's Act promotes children's rights and participation in matters affecting them, including their online safety.[144] The Criminal Law (Sexual Offences and Related Matters) Amendment Act defines "child pornography" as:

> "any image, *however created*, or any description or presentation of a person, *real or simulated*, who is, *or who is depicted or described or presented as being*, under the age of 18 years, of an explicit or sexual nature, whether such image or description or presentation is intended to stimulate erotic or aesthetic feelings or not, including any such image or description of such person—
> (a) engaged in an act that constitutes a sexual offence;
> (b) engaged in an act of sexual penetration;
> (c) engaged in an act of sexual violation;
> (d) engaged in an act of self-masturbation;
> (e) displaying the genital organs of such person in a state of arousal or stimulation;
> (f) unduly displaying the genital organs or anus of such person;
> (g) displaying any form of stimulation of a sexual nature of such person's breasts;
> (h) engaged in sexually suggestive or lewd acts;
> (i) engaged in or as the subject of sadistic or masochistic acts of a sexual nature;
> (j) engaged in any conduct or activity characteristically associated with sexual intercourse;
> (k) showing or describing such person—
> (i) participating in, or assisting or facilitating another person to participate in; or
> (ii) being in the presence of another person who commits or in any other manner being involved in, any act contemplated in paragraphs (a) to (j); or
> (l) showing or describing the body, or parts of the body, of such person in a manner or in circumstances which, within the context, violate or offend the sexual integrity or dignity of that person or any category of persons under 18 or is capable of being used for the purposes of violating or offending the sexual integrity or dignity of that person, any person or group or categories of persons" (own emphasis).[145]

taken or put in place to ensure that children are not exposed to child pornography and pornography." Section 16(1) of the Cybercrimes Act outlines the sharing of intimate images as an offence and states that "Any person (''A'') who unlawfully and intentionally discloses, by means of an electronic communications service, a data message of an intimate image of a person (''B''), without the consent of B, is guilty of an offence." The processing of the special personal information of a child is prohibited in terms of sections 26 and 34 of POPIA unless the provisions of sections 27 and 35 are applicable in which case, subject to section 37, the conditions for the lawful processing of personal information as referred to in Chapter 3 must be complied with. Specifically, section 5(1)(a) to (d) of POPIA highlights that: "A data subject has the right to have his, her or its personal information processed in accordance with the conditions for the lawful processing of personal information". Section 7(1) goes further to note that, "This Act does not apply to the processing of personal information solely for the purpose of journalistic, literary or artistic expression to the extent that such an exclusion is necessary to reconcile, as a matter of public interest, the right to privacy with the right to freedom of expression." Section 34 of POPIA notes that "A responsible party may, subject to section 35, not process personal information concerning a child."

[143] Section 18(3)(a) and section 18(5) of the Film and Publications Act 65 of 1996 (accessible here.)

[144] Section 10 of the Children's Act 25 of 2005 (accessible here.)

[145] Criminal Law (Sexual Offences and Related Matters) Amendment Act 32 of 2007 (accessible here.)

Despite these provisions, challenges remain. The reliance on complaints and investigations may result in delayed responses to rapidly evolving AI-generated content. Additionally, the legislation focuses on internet service providers and individual offenders, potentially overlooking other platforms and technologies where AI-generated CSAM might circulate.

Although these legislative frameworks do not make provision for situations where the AI-generated CSAM is not based on an existing person and therefore technically involves no real child, there is a strong case to adopt this approach in South Africa. Currently, only the United Kingdom has announced its intention to adopt legislation criminalising both the use of AI tools designed to generate CSAM as well as the possession of AI-generated CSAM.[146] While not yet enacted, the legislation would make, "it illegal to possess, create or distribute AI tools designed to generate CSAM" with such an offence being punishable by up to five years in prison if enacted. Further, the law would make, "it illegal for anyone to possess AI 'paedophile manuals' which teach people how to use AI to sexually abuse children" and punish such possession with a custodial sentence of up to three years in prison.[147]

Platform accountability

The proliferation of AI-generated CSAM on social media has led to a troubling normalisation of viewing and creating of harmful content. As such, it has become increasingly necessary to interrogate and regulate social media's role is facilitating the dissemination of AI-generated CSAM. Most platforms have responded accordingly by adopting policies aimed at addressing the dissemination of CSAM generally and AI-generated CSAM specifically. Below, and in no particular order, we outline the relevant policies of some of the most widely used social media and GenAI platforms in South Africa:[148]

X

X has a policy on child safety generally, which includes a prohibition on child sexual exploitation specifically. This policy reads as follows:

> "X has zero tolerance towards any material that features or promotes child sexual exploitation. This may include real media, text, illustrated, or *computer-generated media - including GenAI media*. Regardless of the intent, anyone viewing, sharing, linking, or engaging with any kind of child sexual exploitation material contributes to the re-victimization of the depicted children and puts children at an extreme risk of being harmed. This also applies to content that may further contribute to victimization of children through the promotion or glorification of child sexual exploitation." (own emphasis)[149]

Depending on the violation, users may either be requested to remove their post or users may have their accounts suspended. Where the post depicts or promotes child sexual exploitation, their accounts and content will be reported to the National Center for Missing and Exploited Children (NCMEC).

---

[146] F Brown, SkyNews, *AI tools used to generate child abuse images made illegal in "world leading" move* (accessed in February 2025) (accessible here.)
[147] Id.
[148] See K McInnes, *Meltwater, South African Digital & Social Media Statistics 2024* (accessed in March 2024) (accessible here.)
[149] X Help Center, "Child safety" (May 2024) (accessible here.)

Grok

Grok, owned by xAI, has a user policy that states that the service may not be used to promote or engage in the following illegal conduct: the violation of a person's privacy or their right to publicity, depicting someone's likeness in a pornographic manner, and the sexual exploitation of children among others.[150] It, too, notes that suspected CSAM will be reported to the NCMEC. Curiously, the policy also guides users to not, "...circumvent safeguards unless you are part of an official Red Team or otherwise have our official blessing." The "Red Team" appears to reference safety officers hired by Grok to respond to explicit content.[151] In terms of disinformation, in May 2025, Grok reportedly shared disinformation about a white genocide in South Africa in response to unrelated user prompts.[152] In response, xAI said that this was a result of unauthorised modification.[153] The incident demonstrated the fallible nature of GenAI models and the general lack of transparency around these systems.

Meta

Meta's policy does not allow any content or activity that sexually exploits or endangers children.[154] When they become aware of apparent child exploitation, Meta reports it to the NCMEC.[155] Meta generally removes images of nude children, even if shared with good intentions, to prevent potential abuse and misuse.[156] The platforms also work with external experts, including the Meta Safety Advisory Board, to improve its policies and enforcement around online safety issues, especially concerning children.[157] Meta was reported to have incorporated solutions such as StopNCII and TakeItDown, which assist in detecting and removing non-consensual images from platforms.[158] In 2024, Meta also expanded its existing partnership with the Tech Coalition to include sharing signals about sextortion activity.

Within the context of GenAI, Meta uses visible markers and visible and invisible watermarks and metadata embedded when photorealistic images are created using the Meta AI feature.[159] In February 2024, Meta announced that it was in the process of building tools able to identify invisible markers, thus enabling it to label images from Google, OpenAI, Microsoft, Adobe, Midjourney and Shutterstock.[160] In

[150] xAI, "Acceptable Use Policy" (2 January 2025) (accessible here.)
[151] G Kay and J Newsham, Business Insider, *Elon Musk's xAI is hiring workers to rein in Grok as the chabot spits out NSFW content and racial sluts* (accessed in July 2025) (accessible here.)
[152] D Kerr, "Musk's AI Grok bot rants about 'white genocide' in South Africa in unrelated chats" *The Guardian* (accessed in July 2025) (accessible here.)
[153] D Milmo, "Elon Musk's AI firm blames unauthorised change for chabot's rant about 'white genocide'"*The Guardian* (accessed in July 2025) (accessible here.)
[154] Meta, "Child sexual exploitation, abuse and nudity" (27 December 2024) (accessible here.)
[155] Id.
[156] Above n 154.
[157] Id.
[158] The White House, "White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse" (12 September 2024) (accessed on March 2025) (accessible here); See StopNCII.org (accessible here); See Takeitdown.ncmec.org (accessible here.)
[159] Meta, "Labelling AI-Generated Images on Facebook, Instagram and Threads" (6 February 2024) (accessed in March 2025) (accessible here.)
[160] Id.

April 2024, Meta announced its commitment to building GenAI features responsibly.[161] Notably, Meta announced that it had joined forces with Thorn, All Tech is Human, and other tech companies to prevent the misuse of GenAI tools for child exploitation.[162]

TikTok

TikTok has a general prohibition on CSAM.[163] Importantly, this includes content that is AI-generated.[164] Similarly to X and Meta, TikTok endeavours to take immediate action to remove content containing CSAM, subsequently banning accounts that disseminate CSAM, and submitting reports to the NCMEC.[165] TikTok has become a member of the Coalition for Content Provenance and Authenticity (C2PA) to help support responsible and transparent AI-generated content.[166] Moreover, TikTok has adopted the C2PA's content credentials, which enables the platform's systems to instantly recognise and label AI-generated content.[167] In 2023, TikTok was one of 27 organisations who signed a pledge to tackle AI-generated CSAM.[168]

Pinterest

According to its community guidelines, Pinterest enforces a strict, zero-tolerance policy for any content, including imagery, video, or text, or accounts that might exploit or endanger minors.[169] Its prohibition of CSAM extends to content which sexualises minors in the form of cartoons or anime.[170] The platform also works closely with the NCMEC to combat any activities that exploit or endanger minors.[171]

Snapchat

Snapchat prohibits any activity that involves sexual exploitation or abuse of a minor, including the sharing of CSAM.[172] In its transparency report, Snapchat noted its use of active technology detection tools, such as PhotoDNA robust hash-matching and Google's Child Sexual Abuse Imagery (CSAI) Match to "identify known illegal images and videos" of CSAM.[173] This content is then reported to the NCMEC, which then coordinates with domestic or international law enforcement, as required.

[161] Meta, "Meta Joins Thorn and Industry Partners in New GenAI Principles" (23 April 2024) (accessed in March 2025) (accessible here); Vedantam, *Los Angeles Business Journal,* Thorn Uses AI to Protect Children' (29 July 2024) (accessed in March 2025) (accessible here); See alltechishuman.org (accessible here.)
[162] Id.
[163] TikTok, "Combating child sexual exploitation and abuse" (accessible here.)
[164] Id.
[165] Id.
[166] Id.
[167] Coalition for Content Provenance and Authenticity, "Overview" (accessible here.)
[168] IWF, "AI must be a force for good and not a threat to children" (30 October 2023) (accessible here.)
[169] Pinterest, "Community guidelines" (accessible here.)
[170] Pinterest, "Transparency report" (accessible here.)
[171] Id.
[172] Snapchat, "Community Guidelines – Sexual Content" (updated January 2024) (accessible here.)
[173] Snapchat, "Transparency Report January 1, 2024 – June 30, 2024)" (4 December 2024) (accessible here.)

ChatGPT

ChatGPT's position on safety and age is that the platform is not meant to be used by children under 13 years old.[174] In terms of its usage policy, users may not repurpose or distribute output from the service to harm others, including sexualising children.[175] The policy also notes that apparent CSAM will be reported by the company to the NCMEC. In April 2024, OpenAI published its commitment to child safety, noting that it had partnered with Thorn, a non-profit company focused on defending children from CSAM, and a number of tech companies to implement robust child safety measures in accordance with safety-by-design principles.[176] The companies listed in the announcement are Amazon, Anthropic, Civitai, Google, Meta, Metaphysic, Microsoft, Mistral AI, and Stability AI.

Where does this leave us?

Despite the implementation of policies aimed at addressing AI-generated CSAM on social media platforms, the problem remains persistent and troubling. Reports to the NCMEC have revealed a significant rise in online child sexual exploitation, with instances of AI-generated CSAM becoming increasingly prevalent.[177] Over the past two years, the NCMEC's CyberTipline received more than 7,000 reports of AI-generated child exploitation content.[178] Law enforcement in the United States has faced difficulty in taking action against reports of AI-generated CSAM as there is often a disconnect between the quantity of reports and the quality of the information provided.[179]

Many social media platforms have restricted or entirely removed access to application programming interfaces (APIs) that researchers use to access platform data.[180] This limits the ability to study online behaviour and understand its consequences.[181] Moreover, direct partnerships with social media platforms are usually limited to a select few, primarily in the Global North.[182]

In the midst of this, issues regarding age verification persist. While platforms often claim that age verification is too difficult to implement, this argument falls flat when considering the vast troves of data that they hold per user.[183] It therefore seems implausible to suggest that they cannot then also determine a user's age with better safeguards in place and at a higher degree of accuracy.

The European Union has implemented the Digital Services Act (DSA), which compels online platforms to remove illegal content, prohibits targeted advertising towards content algorithms, and provide greater

[174] OpenAI, "Is ChatGPT safe for all ages?" (accessible here.)
[175] OpenAI, "Usage Policy" (accessible here.)
[176] Id.
[177] K McQue, *The Guardian*, Child sexual abuse content growing online with AI-made images, report says (16 April 2024) (accessed in March 2025) (accessible here.)
[178] National Center for Missing & Exploited Children, "GenAI (GAI)" (accessible here.)
[179] Above n 177.
[180] Parry, "Without access to social media platform data, we risk being left in the dark" (2024) 120(3/4) *S Afr J* (accessible here) at 1.
[181] Id.
[182] Id.
[183] G Radauskas, "All social media apps collect user data but Threads is king" *Cybernews* (22 August 2023) (accessible here.)

transparency into algorithms.[184] Importantly, the Act mandates that states in the EU redesign their systems to ensure a higher level of security for children and compels platforms to reconfigure their services to reduce adverse risks to children's mental health.[185] South African researchers face significant challenges due to restricted access to social media platform data. Without more reliable, legal means of accessing this data, they risk falling behind in research on digital wellbeing and online behaviour contributing to the dissemination of CSAM. Evidence generation of the position in South Africa is critical.

RECOMMENDATIONS

The perpetual technical advancements of GenAI and, in contrast, the slow pace of effective regulation complicates the implementation of best practice principles. There is, however, value in continuing to advocate for tech accountability to align with broader societal values. Below, we propose a number of strategies and recommendations that may be deployed to prevent or mitigate the risks and harms caused to children by GenAI.

Driving localised evidence-generation

A number of the examples and data points referred to throughout this discussion document speak to the Global North. This is because there is presently a gap in localised evidence-generation about children in the broader African context, and in South Africa are engaging with GenAI. While comparative data is useful in highlighting the challenges which may occur on home ground, there are various nuances in digital access and literacy which need to be factored in to inform data-driven approaches across diverse communities. The process of evidence-generation should not fall on civil society or corporate social responsibility initiatives; adequate state funding and support is key.

Prioritising children's participation at in the AI lifecycle

Diverse perspectives in the design, deployment stages of a GenAI model are a means to potentially ensure ethical AI practices.[186] Meaningful children's participation in these processes means the impact of the model on children is consistently taken into account. It has been suggested that children's participation can take on the role of tester, informants, and co-designers (at an age-appropriate level).[187] There is value in adopting a co-creative approach in GenAI design with children as it fosters design justice and enables them to exercise greater power in the innovation of tools that they consume.[188] It goes without saying that children's participation requires robust due diligence processes to ensure that it is implemented within ethical boundaries and does not inadvertently compromise their safety.

Standard testing processes to prevent CSAM

[184] Digital Services Act 2022 (accessible here.)
[185] European Commission, "The Digital Services Act (DSA) - Regulation (EU) 2022/2026" (accessible here.)
[186] P Radanliev, "AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development" (2025) 39 *Applied Artificial Intelligence* (accessible here.)
[187] M Giannakos, M Horns, and M Cukurova, "Learning, design and technology in the age of AI" (2025) 44 *Behaviour & Information Technology* (accessible here.)
[188] S Mathiyazhagan and K La Fors, "Children's right to participation in AI: Exploring transnational co-creative approaches to foster child-inclusive AI policy and practice" (2023) 28 *Information Polity* (accessible here.)

Ofcom, the United Kingdom's communications regulator, has suggested that GenAI platforms engage in abusability testing to bolster safety – this is a process to identify the ways in which a platform may be abused by malicious users.[189] This applies to the content of CSAM and aligns with UNESCO's Recommendation on the Ethics of AI, which calls for rights-based ethical impact assessments to appropriately prevent the risks of AI.[190] Applying a similar system may very well necessitate a comprehensive overhaul of current processes in South Africa to ensure clear action and responses from key stakeholders are better streamlined.

Enforcement of existing cybercrimes laws

At the international law level, it is commendable that UN bodies are moving towards the adoption of instruments on GenAI, including on transnational cooperation. Similarly, in South Africa, the recent publication of the draft National AI Policy Framework is promising. However, while these processes unfold, there are existing laws that may be enforced to influence social norms on safeguarding children's digital rights. For example, the Cybercrimes Act, Criminal Law (Sexual Offences and Related Matters) Amendment Act, and POPIA already include provisions speaking to the emergent risks and concerns which have been highlighted in this discussion document. In South Africa, the High Court's judgment in *KS v AM and Another* sets landmark precedent on the disclosure of non-consensual intimate images and impersonation (although the parties in the matter were not minors).[191] The High Court awarded the plaintiff damages amounting to R3.5 million, a first in South Africa. In the judgment, the judge acknowledged that the occurrence of cybercrimes impacted the plaintiff's physical, emotional and mental health; it impacted her ability to thrive.[192]

Prioritising investment in AI literacy and competency

Broadly, MMA has long advocated for the enhancement of digital and media literacy skills to enable conscious participation and engagement in digital spaces, self-regulation, and empower people to report violations of their safety or rights online. In a country with acute wealth disparities and a digital divide along socio-economic lines, a one-size-fits-all approach is not feasible. Rather, programmes for AI literacy and competency must factor in the various developmental stages, differing levels of more general digital literacy, linguistic and cultural diversity, and the role of parents, teachers, and guardians as non-AI experts.

CONCLUSION

The duality of emergent technology is not novel. Where GenAI enables children to learn, form safe communities, express themselves, and enjoy digital play, it is a wonderful tool. Where it enables the exploitation of children and their image rights, promotes the consumption and dissemination of disinformation, and does not process children's personal data in lawful ways, this poses grave risks that

[189] See Ofcom, "Ofcom calls on tech firms to make online world safer for women and girls" (accessed in July 2025) (accessible here.)
[190] UNESCO, "Recommendation on the Ethics of Artificial Intelligence" (2021) (accessible here.)
[191] See the judgment here. See Power Law Africa's case summary here.
[192] Id judgment at paras 16 and 17.

demand robust and effective responses. This discussion document highlights the meaning of children's digital rights and the internet as an enabler of rights, the emerging risks and concerns associated with children's access to GenAI, the barriers to mitigating and preventing some of the harms arising from GenAI, and corresponding recommendations. As South Africa's discourse on AI continues and legal position begins to take shape, MMA hopes that children's digital safety is prioritised and that children are empowered to be engaged and resilient in an ever-changing digital ecosystem.

MEDIAMONITORING AFRICA