NSPCC

Viewing Generative AI and children's safety in the round

JANUARY 2025

EVERY CHILDHOOD IS WORTH FIGHTING FOR

Contents

Acknowledgements 4
Executive summary 5
Introduction 6
Our approach 9
Risks 10
Why are these risks taking root? 13
Why are these risks likely to spread? 14
Solutions 15
What needs to be done? 19
Appendix 1: Glossary 21
Appendix 2: Methodology 22

Acknowledgements

We are grateful to the Oak Foundation for their support for our work, which has made this research and paper possible. We also extend our sincere thanks to AWO for their invaluable research contributions. Finally, we are deeply appreciative of everyone who participated in the research for their time and expertise.

Executive summary

This report outlines the critical need for robust policies and practices to ensure the safety and protection of children in the context of Generative Artificial Intelligence (Gen AI). As this technology rapidly evolves, it presents both opportunities and significant safety risks to children and young people.

Our research, conducted by AWO, identified seven key safety risks associated with Gen AI: sexual grooming, sexual harassment, bullying, sextortion, child sexual abuse/exploitation material (CSAM/CSEM), harmful content, and harmful ads and recommendations. These risks were conceptualised in a child-centred way, considering how Gen AI outputs can be used maliciously to target children, how children's images can be exploited, the harms from consuming Gen AI content, and the potential for children to create harmful Gen AI content.

This research further identified 27 potential solutions to mitigate these risks. These were then categorised through the Gen AI product cycle: development, release, maintenance, and care. While some of the Gen AI solutions identified are theoretical, others are already being implemented by major AI companies. We know from this research that Gen AI can be made safer; whether through a combination of the solutions that we have identified, or through new solutions developed in the future, we want to see this happen.

We have combined the research with our consultations with young people and information from Childline. This has led us to conclude that the following four urgent actions are needed to mitigate these risks:

1. **Adopt a Duty of Care for Children's Safety.** Gen AI companies must prioritise the safety, protection, and rights of children in the design and development of their products and services. Proper governance of training data, transparency in data usage and allowing audits are essential steps in maintaining this standard, as are conducting risk assessments, consulting with child-safety experts, and making mitigation measures transparent.
2. **Embed a Duty of Care in Legislation.** It is imperative that the Government enacts legislation that places a statutory duty of care on Gen AI companies, ensuring that they are held accountable for the safety of children. This legislation should empower regulators to enforce compliance, thereby providing a robust framework for protecting children from the potential harms of Gen AI.
3. **Place Children at the Heart of Gen AI Decisions.** The needs and experiences of children and young people must be central to the design, development, and deployment of Gen AI technologies. Our consultations with the Voice of Online Youth, a group of 14 young people aged 13–16, revealed critical insights into the risks they perceive and who they believe should be responsible for mitigating these risks. Incorporating children's perspectives ensures that Gen AI solutions are both effective and relevant.
4. **Develop the Research and Evidence Base on Gen AI and Child Safety.** Continuous research is vital to understand the evolving risks posed by Gen AI and to develop effective mitigation strategies. The Government, academia, and relevant regulatory bodies should invest in building capacity to study these risks and support the development of evidence-based policies.

Introduction

The NSPCC is a leading children's charity dedicated to preventing cruelty to children and ensuring their safety. We have been at the forefront of the movement to keep children safe online, continuously adapting to the evolving digital landscape. The NSPCC is working to transform the online world so that it is safe for every child to go online.[1] As part of this commitment, we have explored emerging technologies, including virtual and augmented reality environments, to understand their impact on child safety. This report introduces our understanding and policy position on the safety risks associated with another type of emerging technology: Generative Artificial Intelligence (Gen AI).

The position outlined in this paper relates strictly to Gen AI. Gen AI refers to AI systems that generate or produce new synthesised content in response to user prompts. This technology was first introduced to the general public in late 2022, with the launch of tools like ChatGPT. Gen AI is not simply chatbots, and the content developed by it is not limited to text; these models are also capable of producing images, audio, and even short video clips.

Children are frequently early adopters of new technologies, even though this technology has often not been developed with child users in mind. Ofcom's Online Nation study in 2023 found that teenagers and children in the UK were far more likely that adults to have embraced Gen AI, with 79% of online 13–17-year-olds using Gen AI technology, compared with just 31% of adult users (aged 16 and above).[2]

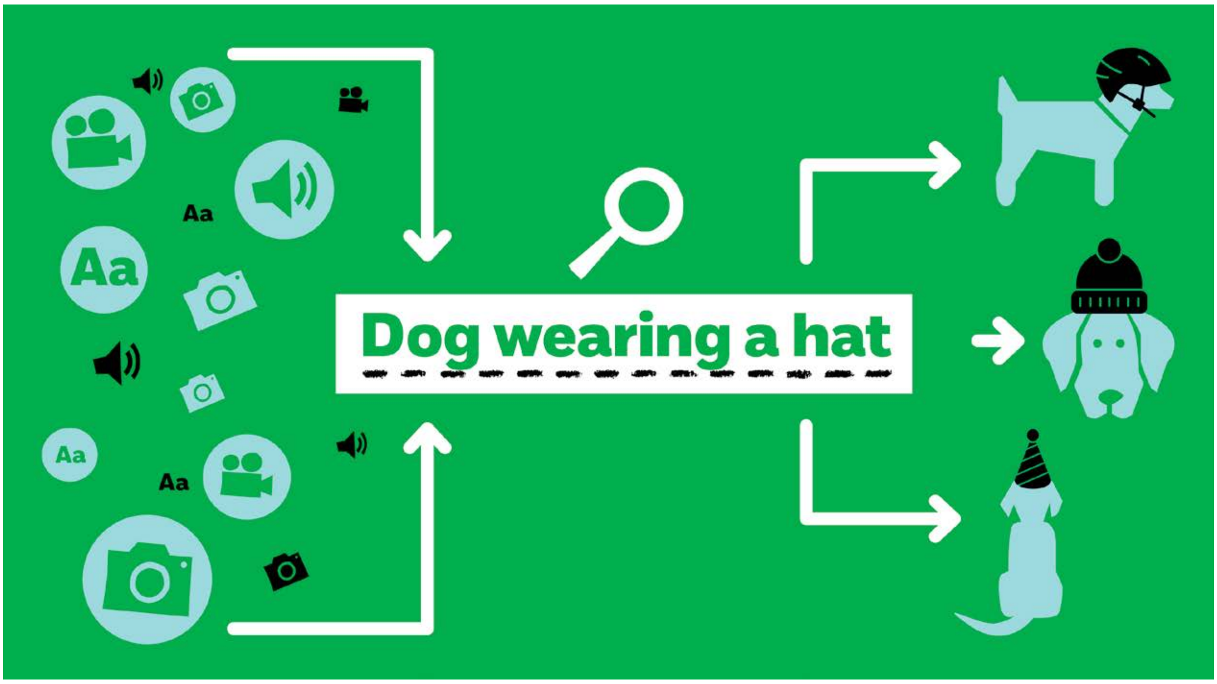

Figure 1: Generative AI models are trained on colossal amounts of data. They use this to generate new content in response to user prompts. So, if, for example, a user prompts the model with 'dog wearing hat', the AI model understands what dogs and hats are, and how to put this knowledge together to create a new image.

[1] NSPCC (2021) *Preventing Child Abuse Together: Our Strategy from 2021 Onwards*. Available online here

[2] Ofcom (2023) *Online Nation 2023 Report*. Available online here

ChatGPT is an advanced AI chatbot developed by OpenAI. It uses a sub-type of Gen AI called a Large Language Model (LLM) to understand and generate human-like text based on the input it receives. This technology works by training the model on a vast amount of text data, scraped from the internet; when you ask ChatGPT a question, it processes your input, predicts the most likely next words, and constructs a response that makes sense in context. This means that ChatGPT can conduct conversations, answer questions, provide recommendations, and help with tasks like writing and brainstorming. Chatbots like ChatGPT can 'hallucinate' and provide false information to users.

New technologies, like Gen AI, come with a number of opportunities for children and young people. Potential benefits of Gen AI include that it can deliver tailored and personalised learning systems for children; these can adapt and evolve to remain customised to a child's changing learning style, to deliver the most impactful possible learning experience. These tailored services can also help to improve learning outcomes for children with disabilities by providing new ways to interact and create in digital systems. Gen AI gives children the opportunity to explore their digital creativity in new and exciting ways. Children can participate in the production of creative content like stories, artwork, games, and other types of software without needing to have expert coding skills.

Gen AI technology is not without its significant drawbacks. As early as 2019, we began receiving contacts from children, via Childline,[3] about harms they have experienced from artificial intelligence. Among mentions of AI in Childline counselling sessions, many involved the generation of sexually exploitative images or videos or the threat to create these alongside blackmail and financial extortion. There were also some examples of young people using chatbots for company, or to help them explore their feelings in a safe space.

> **"I was talking to this girl on Snapchat who I thought was my age, then she said she was actually much older and got angry I didn't want to speak to her anymore. She made fake sexual pictures of me and demanded I send her £200, or she'll send it to my friends. I've reported and blocked the account, but don't know how to be sure they won't send the pictures."**
>
> **Boy, 16, Childline Snapshot**

[3] Please note: The Childline Snapshots provided are based on real Childline service users but are not necessarily direct quotes. All names and potentially identifying details have been changed to protect the identity of the child or young person involved. This allows us to illustrate what we hear from children and young people, without breaching their confidence. This applies to all Snapshots used in this paper.

character.ai is a platform that allows users to create and interact with AI-generated characters. These characters can engage in conversations, provide advice, and play games. The AI learns from interactions, making the characters more responsive and personalised over time – meaning that the characters can remember past interactions with users and adapt their responses accordingly. Characters can converse with users via generated text or audio. The platform has hosted user-generated characters that encourage self-harm, eating disorders, and grooming.[5]

Stable Diffusion is an image generating AI model developed by Stability AI. It uses a technique called diffusion to create high-quality images from text descriptions. Early models of Stable Diffusion focused on basic image generation, while the latest versions offer enhanced customisation, faster processing, and are able to generate video clips and 3D objects. Early versions of Stable Diffusion, without safeguards, can produce sexualised images of children.[6]

The issue of Gen AI safety is important to children; Ofcom's Online Nations study in 2024 found that 46% of internet users in Britain aged 8–15 reported feeling worried about the future impact of Gen AI on people.[4]

While we are concerned about the risks posed by this technology, we recognise that revolutionary technologies like Gen AI will play a significant role in the future lives of today's children and young people, whether they use it directly or not. The solution will not lie in banning children and young people from this technology, but instead in implementing safety measures to protect them.

[4] Ofcom (2024) *Online Nation 2024 Report*. Available online here

[5] Maggie Harrison Dupré (2024) *AI Chatbots are Encouraging Teens to Engage in Self-Harm*. Futurism. Available online here
Maggie Harrison Dupré (2024) *Character.AI Is Hosting Pedophile Chatbots That Groom Users Who Say They're Underage.* Futurism. Available online here
Maggie Harrison Dupré (2024) *Character.AI Is Hosting Pro-Anorexia Chatbots That Encourage Young People to Engage in Disordered Eating. Futurism.* Available online here

[6] Angus Crawford and Tony Smith (2023) *Illegal trade in AI child sex abuse images exposed.* BBC News. Available online here

Our approach

This paper combines analysis from research commissioned by the NSPCC with publicly available data and the views of children and young people to outline the current risks to children's safety posed by Gen AI. It outlines the potential solutions to these risks and the necessary policy response.

We commissioned AWO, a legal and technology consultancy, to consult experts from a wide range of sectors and identify evidenced and hypothetical risks to children's safety and how they may be mitigated.[7] A panel of 11 young people aged 13–16 from the NSPCC's Voice of Online Youth were asked to give their perspectives on Gen AI risks and who they felt was responsible for addressing these risks. We also gathered relevant insights from Childline, ensuring children's voices were central to our policy development. The results of this work were then sense-checked with stakeholders.

Overall, we concluded that a multipronged approach should be taken to embed children's safety in Gen AI. Companies must take children's safety seriously, and child protection must be central to technical and legislative changes – changes that should be informed by children's voices. Further investment is needed in research in order to ensure that companies and governments have the information they need to keep children safe. These measures are expanded on later in this report in 'What needs to be done?' (page 19).

[7] For further detail of our methodology, please see Appendix 2.

Risks

While the public discourse has often focused on the threat of AI-generated Child Sexual Abuse Material (CSAM), the research identified evidence of a total of seven safety risks associated with Gen AI. These are:

- sexual grooming
- sexual harassment
- bullying
- sexual extortion
- child sexual abuse/exploitation material (CSAM/CSEM)
- harmful content
- harmful ads and recommendations

The risks associated with Gen AI were conceptualised in a child-centred way, whereby children could be the target of Gen AI, the subject of Gen AI, the consumer of Gen AI, or its creator. There was no attempt during the research to measure the extent or scale of these risks, which remain unknown.

"I'm still so anxious about what happened at school last year. These boys made fake porn of the girls, including me, and sent them to loads of group chats. They were excluded for a bit, and we had a big assembly about why it was wrong, but after that school told us to forget what happened. I can't forget though, people think that they saw me naked, and I have to see these boys every day."

Girl, 14, Childline Snapshot

Target: how Gen AI outputs can be used maliciously to target children

"A group of boys at school used deepfake to make a video of me saying I'm gay. They've made fake chat screenshots of me saying I want to do sexual things to them as well. I have questioned my sexuality but haven't come out to anyone, that doesn't stop the bullies though. I want to tell a teacher but it's my word against all these other boys."

Boy, 14, Childline Snapshot

Children can be placed at risk by being deliberately targeted with the outputs of Gen AI. The creation and deployment of AI-generated voice and image outputs can facilitate a range of harms. Malicious actors can use image generators to de-age their photos and change their gender in order to create fake child profiles; these, in turn, can be used to approach or connect to children online for the purposes of grooming, sexual harassment or sextortion. Adults and children can create deepfake personas using voice and video generation technology, and use these to present themselves as potential friends or romantic partners to prospective victims (similar to how real-time deepfake technology is used in adult romance scams).[8] Deepfakes of children saying or doing something that they did not do in real life, or where their physical appearance has been

[8] Matt Burgess (2024) *The Real-Time Deepfake Romance Scams Have Arrived*. WIRED. Available online here

altered, may be used to target them with humiliation and non-sexual bullying, while Gen AI chatbots can additionally serve to provide suggestions of how a victim could be bullied. Sexual deepfakes of children (CSAM or CSEM) may be used to threaten or blackmail them into doing something they do not want to do.

In addition, Gen AI models that collect personal or behavioural information about their users can exploit this information to algorithmically target children with harmful ads and harmful recommended content. For example, children who have inadvertently revealed they have concerns about money might be shown gambling ads, or those interacting with "AI girlfriends" might be exposed to misogynistic content.

Subject: how children's images can be assimilated into harmful Gen AI outputs

Children can be victimised by Gen AI tools that exploit their image to create sexually explicit images of them. These images can either meet the criminal threshold (CSAM) or not (CSEM).[9,10,11] Such images can be created deliberately using face-swapping and nudifying Gen AI models, or by artificially placing existing images of children in sexual contexts, poses and situations. Gen AI models that have been trained on CSAM or pornographic images can also create new CSAM/CSEM, inadvertently or otherwise. Perpetrators share manuals explaining how these images can be created and finessed.

The malicious creation of non-sexual images of children through Gen AI can be equally harmful. Gen AI can be used to embarrass a child by placing their image in humiliating contexts or positions, or manipulate an image to depict a child self-harming. These uses can have a significant negative emotional impact on victims.

> **"I'm so ashamed of what I've done, I didn't mean for it to go this far. A girl I was talking to was asking for pictures and I didn't want to share my true identity, so I sent a picture of my friend's face on an AI body. Now she's put that face on a naked body and is saying she'll post it online if I don't pay her £50. I don't even have a way to send money online, I can't tell my parents, I don't know what to do."**
>
> **Boy, 14, Childline Snapshot**

[9] Thorn (2024) *Youth Perspectives on Online Safety, 2023*. Available online here

[10] National Center for Missing and Exploited Children (2024) *Generative AI CSAM is CSAM*. Available online here

[11] Internet Watch Foundation (2023) *How AI is being abused to create child sexual abuse imagery*. Available online here
Internet Watch Foundation (2024) *What has changed in the AI CSAM landscape?* Available online here

Consumer: how children can be harmed by consuming Gen AI content

Children's physical and mental health, their development, and their perception of society can be placed at risk through viewing or interacting with Gen AI content. For example, they may be distressed by exposure to AI-generated CSAM or CSEM or by coming across violent or disturbing images and messages created by AI. They may consume AI-generated disinformation or misinformation in their social feeds. Or they may interact with chatbots that bully them, give them false medical advice, or steer them towards eating disorders or self-harm.[12]

> **"Can I ask questions about ChatGPT? Like how accurate is it? I was having a conversation with it and asking questions, and it told me I might have anxiety or depression. It's made me start thinking that I might?"**
>
> **Girl, 12, Childline Snapshot**

Creator: how children who create Gen AI content can harm other children

Children can use Gen AI to harm others in all the ways described above. They can use Gen AI tools to create CSAM and CSEM; to create new content that they then use to bully, sexually harass or blackmail their peers; or to create content that misinforms their peers or encourages them to self-harm. They may do this without fully realising the gravity of what they have done – for example, by creating an image of a friend intended as a joke, but that deeply upsets the victim.

[12] Maggie Harrison Dupré (2024) *AI Chatbots are Encouraging Teens to Engage in Self-Harm*. Futurism. Available online here
Maggie Harrison Dupré (2024) *Character.AI Is Hosting Pedophile Chatbots That Groom Users Who Say They're Underage*. Futurism. Available online here
Maggie Harrison Dupré (2024) *Character.AI Is Hosting Pro-Anorexia Chatbots That Encourage Young People to Engage in Disordered Eating*. Futurism. Available online here

Why are these risks taking root?

The types of social factors that facilitate the victimisation of children online can also heighten children's vulnerability to the risks associated with Gen AI. Low digital literacy, the culture of shame that so frequently surrounds abuse, and children's individual vulnerabilities can increase the likelihood of victimisation. Risks are further enhanced by an online environment that encourages social networking, and in which child images are readily available for perpetrators to exploit.

"I'm not in a good place right now. I don't want to bother others with my problems, so I use ChatGPT to vent and talk about my feelings. I get some validation but, in the end, I still feel lonely."

Young Person, 17, Childline Snapshot

Technological factors can further facilitate these risks. The ease with which AI content can be generated; the availability of open-source Gen AI models (see box) and the integration of Gen AI into other, more readily accessible services; encryption; and ineffective age gating – among other factors – remove barriers to perpetration. The risk to children is reinforced by the realism of Gen-AI outputs, their relatability to children (see box) and their incorporation in training data (see box).

Anthropomorphisation refers to the attribution of human traits to Gen AI models and systems. This happens when Gen AI systems produce human-like responses or are marketed as "digital friends". The perceived humanity of these systems means that children are more likely to trust the information that the system provides to them and are more at ease sharing highly sensitive information with chatbots.

The presence of **CSAM in training data** significantly contributes to the ability of a given model to generate new CSAM. Image generating models can adapt existing CSAM to produce new images of a child who has been sexually abused, placing them in different poses and contexts. Face-swapping technology can be used to edit images of children who have not been victimised. Gen AI models, trained on vast internet-sourced data, may unknowingly include CSAM and, therefore, be capable of generating more, as seen with the LAION-5B dataset used for models including Stable Diffusion. Models trained on images of children alongside adult sexual content can also generate CSAM.

The majority of AI-CSAM and AI-CSEM is currently generated by **open-source models**. Closed, commercial models have safeguards in place that make it difficult to create CSAM. In contrast, early open-source models did not have these safeguards in place, and it is possible for a technically advanced perpetrator to remove safeguards from current open-source models. They can then further train these models to make them better able to create CSAM. Once an open-source model has been downloaded, it is almost impossible to track. Alongside the creation of CSAM, open-source models are also able to create other forms of harmful content, such as disinformation, without this being monitored by developers.

Why are these risks likely to spread?

As this technology develops, so too will the risks. The increasing capability of video-generation models, for example, creates the potential for greater quantities of increasingly realistic CSAM. Without appropriate safeguards, perpetrators can use these advanced video generating models to target children with sexual harassment, bullying, grooming, and sextortion, or with misinformation and disinformation.

As Gen AI outputs become more widespread and realistic, and the models used to produce them become increasingly accessible and integrated into child-facing services, children will correspondingly become more exposed to the risks they pose. Distinguishing between real and AI-generated content will likely become more difficult, which may increase the impact of many of the risks described above. As Gen AI starts to be combined with other emerging technologies, such as augmented or virtual reality, the number and types of environments in which perpetrators can target children will expand.

Children are already experiencing a wide variety of complex harms from Gen AI. Without rigorous safety measures in place, these risks are likely to continue to proliferate and evolve, and new risks will emerge. It is vital that measures are taken to protect children from these harms.

Solutions

The research we commissioned identified 27 potential solutions that can help prevent or mitigate the risks outlined above. These were classified according to which stage in the life cycle of Gen AI development they belong and the risks they target. A legal, technical, and cultural lens was applied to assess how readily these solutions could be implemented.

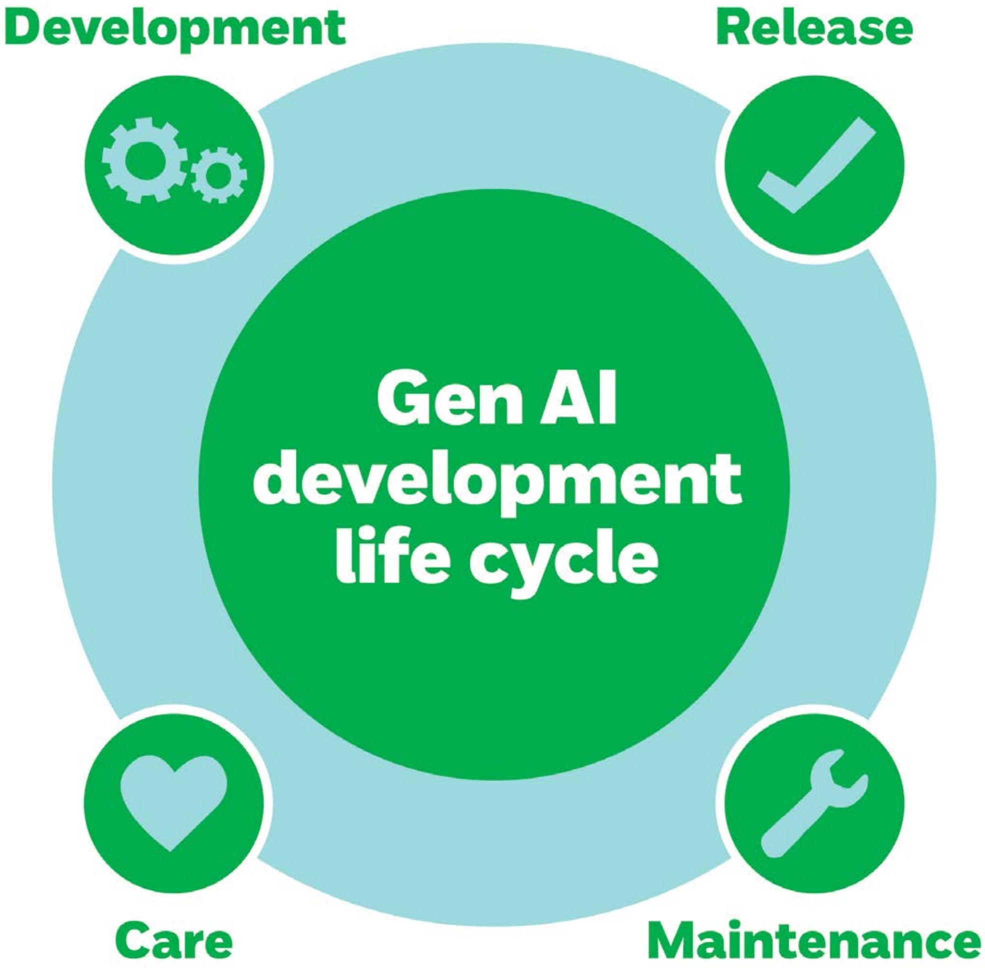

Figure 2: The life cycle of Gen AI development

The research additionally explored whether there were ways in which Gen AI was being used to help safeguard children online. It highlighted instances where GenAI has contributed to content moderation and law enforcement, prevention messaging, trauma counselling, and the detection of child sexual abusers. It is important to note that the use of Gen AI to protect children is largely theoretical at this stage. While there are AI solutions to protect children currently in use,[13] solutions that specifically use Gen AI are in their infancy and represent a potential area for future development.

[13] Neil Sahota (2024) *AI Shields Kids By Revolutionizing Child Safety and Online Protection*. Forbes. Available online here

Gen AI is a technology that is rapidly advancing. The solutions that have been identified represent what is currently available or will soon be available. Rather than being aspirational, these solutions are a baseline; we want to see more, and better solutions being developed and implemented in the future. Indeed, companies including Meta, Google, and Microsoft have already committed to implement some of the solutions identified, as part of their involvement in Thorn's *Safety by Design for Generative AI work.*[14]

Below is a small sample of the solutions suggested at each point in the Gen AI life cycle.

Development: Standards for model creation and training

Training data governance

- Training data must not contain CSAM, CSEM, nude depictions of children, or any depictions of children when also training on adult sexual content.
- Developers should be transparent about their training data and allow it to be audited.

Prevention messaging

- Models or interfaces should include safety warnings for prompts that could return problematic content.
- CSAM prevention Chatbots should be implemented, as proposed by the Lucy Faithfull Foundation.[15]
- Government should raise awareness in developer communities about the potential harmful uses of Gen AI and the necessary safeguards.

Risk assessments

- Risk assessments should be conducted on models to ensure data governance and that safeguards are in place.
- Children and child-safety experts should be consulted during these assessments.
- Risks must be assessed before and continuously after model release.
- There should be transparency around model evaluations and mitigating measures.

> **"A lot of the problems with Generative AI could potentially be solved if the information [that] tech companies and inventors give [to] the Gen AI was filtered and known to be correct."**
>
> **Voice of Online Youth Participant**

> **"Tech companies need to be held responsible for allowing their AIs to generate illegal content and allowing their AI to use illegal data for its model."**
>
> **Voice of Online Youth Participant**

[14] Thorn (2024) *Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments*. Available online here

[15] Sarah Smith (2024) *Chatbots and warning messages: innovations in the fight against online child sexual abuse*. The Faithfull Papers. Available online here

Release: Responsible release standards

Child safety-focused detection and content moderation tools

- Child safety-focused detection and content moderation tools should be deployed to detect and reject harmful prompts.
- Filtering must be adapted continuously to address alternative language use.
- Users who repeatedly attempt to create CSAM should be banned and potentially flagged to law enforcement.
- Users must be given the ability to flag and report harmful content.

Open-source release standards

- Minimum standards should be established for open-source releases to hold developers accountable and ensure implementation of safety measures.

Maintenance: Online services due diligence

Age assurance

- Effective age assurance techniques must be used to prevent children from accessing technology that is not appropriate for their age group.
- Highly effective age assurance should be implemented on social media platforms to safeguard children's accounts.

> **"Social media companies NEED to be able to interpret and display content warnings based on Gen AI content."**
>
> **Voice of Online Youth Participant**

Provenance marking

- AI-generated content on social media and messaging services must be labelled, for example through the use of machine-readable watermarks.
- Labelling should be used to facilitate content moderation and law enforcement efforts.

Care:
Education and care guidance

> **"Everyone has to become more aware about Gen AI and its harmful impact."**
>
> **Voice of Online Youth Participant**

- Guidance should be provided to support parents and teachers to help safeguard children from the risks of GenAI. These should include:
  - Resources that demystify Gen AI and explain its use of data, shortcomings, and risks.
  - Conversation guides for discussing Gen AI risks with children.
  - Experiential learning guides.
  - Tips to encourage critical thinking in children.
  - Advice for identifying behavioural changes that may indicate a child is a victim of sextortion, bullying, or harassment.
  - Guidance on whether, where and when to publish images of their children online.

This is just a taster of the solutions currently available; the full solutions framework can be made available on request.

We know from our work that Gen AI can be made safer; whether through a combination of the solutions that we have identified, or through new solutions developed in the future, we want to see this happen.

What needs to be done?

We propose that tackling the risks and taking advantage of the solutions posed by Gen AI is approached in a principles-based and technology-agnostic manner. This will allow adaptation to new risks and new solutions as the technology develops and the situation evolves.

Measures need to be taken to safeguard children from the risks posed by Gen AI, while also preserving their access to this technology.

There is no one actor with sole responsibility for safeguarding children from Gen AI risks. From developers at the earliest stages ensuring that training data is properly governed, to consumers ensuring that they use Gen AI tools responsibly, everyone is going to have a role to play in the whole-system response that is needed. Child safety needs to be embedded at every stage of the Gen AI development and consumption pipeline.

To achieve this, the NSPCC recommends that the following four actions are taken urgently:

1. Companies must adopt a duty of care for children's safety

Companies must put safety, protection, and the rights of children at the centre of efforts to design and develop products and services. The solutions identified by the NSPCC's research are a good starting point for companies looking to address children's safety in their work.

In particular, companies should be focusing on risk assessing their products, understanding which risks children are likely to be exposed to by their products, and which solutions will be the best fit for mitigating these. Companies should also be using this as an opportunity to identify where new solutions are needed, where current solutions could be pushed further, and innovate to ensure that their products are safe. Companies, both in the UK and abroad, can collaborate on best practices for risk assessing and on the development of new solutions.

2. The UK Government must ensure that legislation embeds this duty of care and enables regulation of Gen AI

While we support current efforts being made to get technology companies on board with voluntary safety commitments, we need government to go further. We know, from our experience campaigning for online safety measures, that self-regulation of tech companies is flawed and that legislative measures will be necessary to bring about meaningful change. To ensure services embed child protection, the UK Government needs to place companies' duty of care on a statutory footing and empower relevant regulatory bodies to enforce this. It additionally needs to ensure that child protection is a central element of its strategic work around AI and of any future AI legislation, including the upcoming AI Bill.

3. Decision-makers must put children's views and experiences at the heart of action on Gen AI

Children and young people's needs and experiences must inform Gen AI design, development and deployment. Globally, government, regulators, and companies must all be engaging with children and young people to understand their views on this technology, which will shape their lives. Children have particularly valuable input to give when it comes to developing educational resources and guidance about the safe use of Gen AI.

4. Everyone should promote the development of the research and evidence base on Gen AI and child safety

The research we commissioned revealed numerous gaps in the evidence base and made suggestions for future projects. At present, there is little understanding of the scale at which many of the risks we found are occurring, or their specific impacts on young people. Data on the experiences of children from diverse backgrounds is particularly lacking.

Furthermore, while our experts were able to propose 27 solutions, more work needs to be done to fully assess what the impact of these would be. While the NSPCC's research in this area has broken new ground, we cannot do this alone. Governments, academics, and regulators need to build capacity – at pace – to better understand risks and mitigations of this rapidly evolving technology.

Appendix 1: Glossary

Anthropomorphisation: When Gen AI models and systems are attributed human traits, such as when they are marketed as "intelligent assistants" or "digital friends", when they produce human-like responses, or when they are given human names.

Age assurance: An age gating technique that involves determining that a user is above the required age, without necessarily needing to determine their exact age.

Chatbot: Chatbots, or "conversational AI" allow the users to "speak" with Gen AI models using everyday language. Chatbots interpret instructions and questions and provide responses.

CSAM: Child Sexual Abuse Material (CSAM), i.e. content that depicts acts of child sexual abuse and/or which focuses on the genitalia of children.

CSEM: Child Sexual Exploitation Material (CSEM), i.e. material that is sexually exploitative of children (for example, revealing clothes, suggestive poses), but does not necessarily meet the definition of CSAM.

Deepfake: Deepfakes are images or videos that have been digitally altered so they appear to show someone else, typically for malicious purposes. Deepfakes can be non-consensual intimate images (NCII), i.e. sexually explicit deepfakes created without the consent of the subject.

Face-swapping: Swapping faces between images or in a video, while maintaining the rest of the body and context.

Gen AI: Generative AI (Gen AI) is a field of AI that focuses on creating new content based on existing data. Inputs and outputs of Gen AI include text, images, video and voice content.

Gen AI companies: Companies that develop Gen AI models and systems, for example Meta, Google, OpenAI, Stable Diffusion, Microsoft and Snap Inc.

Open-source models: "Open-source" is here used loosely to refer to models whose components have been openly released to the public in a way that allows anyone to use, adapt and build on the model.

Safeguards: Model safeguards are tools and structures put in place by developers that help ensure the model behaves in the intended manner. They can include, for example, input filters (preventing the use of certain inputs), watermarks and data governance practices.

Prompt: A prompt is the input that users give Gen AI models. They can include commands ("summarise this for me") and questions ("what is…?") provided through text or voice and can be accompanied with other forms of media ("change this image").

Training data: The data used to train a Gen AI model. This data is typically scraped from the internet and social media platforms. It may include text, images, video and sound, and can include content representing children, and children's data. Some training data sets are also available on an open-source basis for everyone to use.

Appendix 2: Methodology

Commissioned Research

AWO carried out research to establish the risks that Gen AI poses to children's safety; and to explore how these risks are, or could be, mitigated or prevented through technology, legislation or regulation. A secondary objective was to identify any ideas and products in development or currently in use that deploy Gen AI to safeguard children.

The research consisted of 34 qualitative interviews with professionals who have expertise in Gen AI and/or child safety, followed by a workshop to refine the findings. Fieldwork was conducted between May and September 2024. Participants included representatives from child safety and digital rights NGOs, the technology industry, 'safety tech' providers, Gen AI developers, academia, law enforcement, and policymakers.

Participatory session

The Voice of Online Youth is a group of 14 volunteers aged 13–17 whose role in the NSPCC is to ensure that young people's perspectives on online safety are heard and factored into decision-making.

NSPCC staff conducted a one-hour participation session with 11 members of the group in August 2024. Participants were asked to brainstorm the risks associated with Gen AI and, using risky scenarios as prompts, discuss who they felt was responsible for the prevention and mitigation of those risks. It is important to note that the Voice of Online Youth is a small group of highly engaged young people whose views may not be representative of young people more generally.

NSPCC

Together we can help children who've been abused to rebuild their lives. Together we can protect children at risk. And, together, we can find the best ways of preventing child abuse from ever happening

We change the law. We visit schools across the country, helping children understand what abuse is. And, through our Childline service, we give young people a voice when no one else will listen.

But all this is only possible with your support. Every pound you raise, every petition you sign, every minute of your time, will help make sure we can fight for every childhood.

nspcc.org.uk

Photography by Tom Hull. The adults pictured are models and volunteers. J20240003