DECEMBER 2024

Evolving Technologies Horizon Scan

A review of technologies carrying notable risk and opportunity in the fight against technology-facilitated child sexual exploitation.

A joint initiative of Thorn and WeProtect Global Alliance

THORN

weprotect Global Alliance

Table of Contents

Acknowledgments

Understanding the complex intersection of technology and child sexual abuse allows us to prevent and respond to the evolving threats children face online. This can be done only through multistakeholder collaboration that builds on the insights and talents of all in the ecosystem: nonprofits, policymakers, tech companies, communities, investigators, and — most importantly — the young people we commit to serve.

OUR THANKS

We are grateful to the individuals who took the time to participate in this research. Without their participation, we could not share the valuable insights included in this report. In addition to the research and design teams below, we are grateful for the guidance and input of our steering committee members, whose global expertise served as a vital input to delivering this unique look at evolving and emerging threats to online child safety.

Steering committee membership

Participation in the steering committee does not imply endorsement (in part or full) of the findings presented in this report.

Adele Desirs
Queensland Police Service

Campbell Wilson
Monash University

Dan Sexton
Internet Watch Foundation

David Thiel
Stanford Internet Observatory

Fred Langford
Ofcom

Hany Farid
University of California, Berkeley

John Shehan
National Center for Missing & Exploited Children

Katherine King
eSafety Commissioner's Office

Nina Vaaranen-Valkonen
Protect Children

Richard Chambers
International Criminal Police Organization

Sean Litton
Tech Coalition

Sonia Livingstone
London School of Economics and Political Science

This report is a joint initiative between Thorn and WeProtect Global Alliance.

WeProtect Global Alliance brings together over 300 members from governments, the private sector, civil society, and intergovernmental organizations to develop policies and solutions to protect children from sexual exploitation and abuse online. For more information, please visit www.weprotect.org.

Thorn is a 501c(3) organization with a mission to build technology to defend children from sexual abuse. For more information about Thorn, please visit our website: www.thorn.org.

For inquiries about this project, please email research@thorn.org or info@weprotectga.org.

Research team:
Melissa Stroebel, Thorn
Dr. Rebecca Portnoff, Thorn
Caroline Neiswender, Thorn
Shailey Hingorani, WeProtect Global Alliance
Iain Drennan, WeProtect Global Alliance

Design and publication:
Yena Lee, Thorn
Cassie Coccaro, Thorn
Michelle Jeuken, WeProtect Global Alliance
Karen Mulvee, AudienceNet
Loan Nguyen, AudienceNet

Suggested citation:
Thorn and WeProtect Global Alliance. (2024). Evolving Technologies Horizon Scan: *A review of technologies carrying notable risk and opportunity in the fight against technology-facilitated child sexual exploitation.*

Overview

Adoption of new technologies is accelerating, with increasing access around the world, including by children and young people. Hundreds of millions of children around the world use the internet,*[1,2] exploring, creating, learning, and playing. While the benefits of technology are undeniable, society has failed to adequately anticipate and address unintended consequences with regards to child safety before widespread adoption began.

Prior to widespread adoption of the internet, those who sought to sexually exploit and abuse minors faced significant barriers for access to victims, abuse imagery, and other like-minded individuals. Child sexual abuse material (or CSAM) was printed or shipped by mail on VHS tapes and other media formats. Access was slow, with significant hurdles.

However, the internet lowered many of those hurdles. CSAM could be located via the search bar, victims could be found in chat rooms, and entire communities of users could exchange tactics and homemade abuse imagery in seconds. The scale of child sexual abuse and exploitation online skyrocketed, technology evolved quickly, and adequate guardrails have been lacking.

Too often we are failing to anticipate the next priority in online child safety. Taking a reactive approach leaves us grappling with high-stakes consequences and abuse that has already occurred, rather than proactively addressing the potential for harm and preventing this from occurring in the first place.

We need a paradigm shift in how we develop and deploy technology, with builders and protectors collaborating to successfully prevent harm without stifling innovation.

WeProtect Global Alliance and Thorn have teamed up with a community of global experts to support a different approach.

Collaborative horizon scanning is a strategic process that involves systematically gathering and analyzing information to identify emerging trends, issues, and opportunities that could have significant impacts in the future. A proactive approach, it relies on initial or early signs or indicators that suggest the emergence of a new trend or development. It allows early reflection on these developments' intended and unintended effects — the game changers that could make a significant societal and policy impact.

In 2024, the organizations launched the Global Evolving Technologies Horizon Scan project, an initiative designed to examine the most critical technology trends whose evolution are and will significantly impact the fight against technology-facilitated child sexual exploitation. This report is a snapshot of these technology trends, recognizing their value while identifying ways in which the technologies are being (or stand a high risk of being) abused in service of child sexual exploitation. It also offers an initial review of literature exploring safety by design[3,4] tactics for building these technologies and the surrounding platforms to better anticipate, reduce, prevent, and combat abuses online.

*This number is derived from data showing that roughly 33% of children have internet access in the home (see endnote 1) and that roughly 2.2 billion people are under 18 globally (see endnote 2). This is likely a low estimate of 2025 numbers.

The following technologies are discussed in this report: predictive artificial intelligence, generative artificial intelligence, end-to-end encryption, extended reality environments, decentralization, and quantum computing. While the individual technologies were found to present distinct risks and opportunities, requiring differing approaches to mitigating abuse, three core themes emerged across the collection of technologies explored.

There is no one singular technology that we can focus on at the expense of others.

The technologies explored in this report do not exist in silos, nor do kids (or offenders) use them in isolation. Failure to consider multiple technologies in parallel will result in gaps in protection and unattended vulnerability.

A child's privacy is part of their safety — and both must be championed.

A false narrative has dominated some debates around the best ways to safeguard children from risks of online abuse and exploitation — one end of the spectrum calls for removal of privacy and access to content through monitoring and technology bans, and the other relies solely on reacting to disclosure of abuse from impacted children and reports from the public, after the harm has occurred. Neither end of the spectrum will do justice to children's safety and rights.

Technology will be part of the solution to technology — but it will require specialized expertise for efficacy and public/market will.

While technology has exacerbated and accelerated risks of sexual exploitation online, it also is critical in combating these online harms. Building the right solutions, however, will require diverse and specialized expertise across sectors and experience. More, prioritizing the investment in this work will require both public and market will.

This report is a first step in engaging colleagues in continuous, cross-cutting reflection around emerging and potential future trends and starting to consider how we might equip practitioners and experts with the tools to anticipate and tackle the ever-evolving challenges and opportunities children face in the digital world. As such efforts continue, further investment in a diverse portfolio of research initiatives that leverages the knowledge of technologists, investigators, trust and safety professionals, and other frontline experts stands to deliver significant returns. Critically, engaging with and learning from young people must be a central component of these efforts. Building with their perspectives and needs at the forefront will be vital to ensuring the next generation has access to the safe and vibrant digital world they deserve.

Methodology & Design

This study pursued a collaborative horizon-scanning approach, seeking to assemble and learn across a multitude of signals to examine evolving and emerging trends anticipated within the next 5-10 years.

Challenges & Limitations

CHALLENGE: Different global contexts and levels of technology adoption.

The landscape of technology-facilitated child sexual exploitation looks different around the world, with the role of technology and its level of adoption impacting the nature of the harm. Legal and cultural standards differ greatly globally, resulting in differences such as the definition of a child or what constitutes child sexual exploitation or abuse, as well as the presence of social support and protections for the impacted victims. Where some parts of the world are highly integrated with technology, others do not yet have widespread access. While this report sought global input and was framed within the context of international human rights standards, the majority of survey and steering committee participants were based in the global north. This report should not be viewed as fully representative of the unique experiences relating to technology-facilitated child sexual exploitation impacting minors and their communities in the global south.

CHALLENGE: Published literature on the nature of technology-facilitated risks is limited.

The velocity of technological innovation is often mismatched with that of academic literature publication. In addition, our ability to quantify and describe whether and how a technology is, in fact, impacting society lags behind its deployment. Finally, the ability to research the specific intersections between a technology and child sexual exploitation can be hampered by the sensitive and, at times, illegal nature of relevant data. Significant privacy considerations are in play, and access to child sexual abuse material, specifically, is prohibited in most non–law enforcement settings. This report looks across a variety of source types in an attempt to somewhat mitigate this. The research draws from academic literature and grey literature, as well as primary accounts from those with direct experience on this topic. This research did not include any direct analysis of abusive chats or child sexual abuse material.

CHALLENGE: Silos across areas of expertise and perspectives inhibit our ability to forecast, identify, and build to mitigate the misuse of technologies for online child sexual exploitation.

Understanding the impact of technologies on child sexual exploitation requires examination of multiple fronts: technical capabilities, offender tactics, and child development, to name only a few. These areas of expertise are often siloed in daily

business if operational leads do not prioritize intentional, proactive collaboration. That collaboration is necessary to best anticipate and respond to emerging threats relating to technology-facilitated child sexual exploitation. In addition, young people's experiences and perspectives must be at the heart of our work. Engaging with minor audiences requires specialized care and planning to ensure their experience is safe and supportive. To address some of these needs, this research identified expertise across several sectors — including academia, technology, law enforcement, online regulators, civil society, and child development — to represent distinct perspectives and expertise in this research. This research did not engage directly with minors and was limited to the experiences and perspectives of those adults who have worked on their behalf.

CHALLENGE: Technology and social norms are constantly evolving, requiring a commitment to recurring evaluation of their intersection with technology-facilitated child sexual exploitation.

The capabilities of technologies in this report are changing as we speak. So, too, are the social norms and policies surrounding their usage. The risks and opportunities discussed here reflect the current state of this issue, at the time of report production. As a result, this report is intended to provide a snapshot at this moment in time, and best practice will necessitate reevaluation of harms and opportunities on an ongoing basis.

Design

In light of online child safety's complex and dynamic nature, this project took a multipronged approach to balance completeness and velocity. This work built off the existing subject matter expertise of Thorn and WeProtect Global Alliance via three primary workstreams:

- **Workstream 1:** Steering Committee Consultation
- **Workstream 2:** Ecosystem Survey
- **Workstream 3:** Literature Review

Workstream 1: Steering Committee Consultation

The workstream focused on collaboration and learning with the 12 field experts participating in the project steering committee. This was done through survey and subsequent group discussions.

SURVEY

A brief online survey was prepared and fielded among the 12-member steering committee. The purpose of this survey was to:

1. Collect a list of core technologies that increase the risk to children of online sexual exploitation to be evaluated in this initiative.
2. Collect a list of core technologies presenting opportunities to safeguard children from online child sexual exploitation to be evaluated in this initiative.

DISCUSSION

The findings of this survey were used to discuss these technologies; they focused on the following core questions:

1. Does the current list of technologies for exploration represent the most pressing to the child safety ecosystem?
2. What particular risks or opportunities do we see relating to these technologies?
3. Are there any additional emerging technologies not yet being explored in this initiative?

Workstream 2: Ecosystem Survey

This initiative focused on hearing from a broader cohort of ecosystem voices. The online ecosystem survey was distributed via organizer and steering committee channels, including membership lists and contacts. The survey was open for 3 weeks and was completed by 280 respondents (Participant sector breakdown available in Fig 1).

The core questions explored in the survey included:

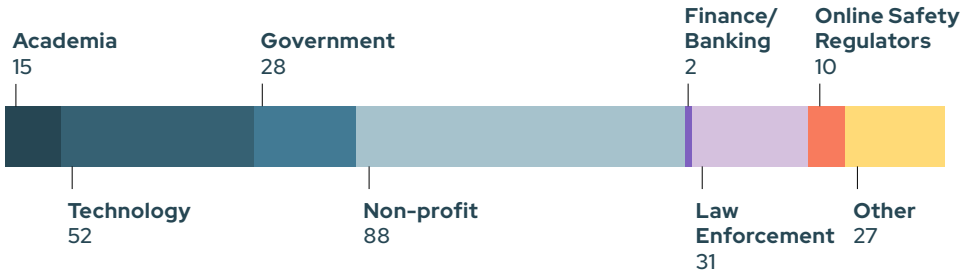
1. Of a given list of technologies, which contribute the greatest current harm relating to online child sexual exploitation and why?
2. Of a given list of technologies, which pose the greatest future risk relating to online child sexual exploitation and why?
3. Of a given list of technologies, which offer the greatest current benefit to combating online child sexual exploitation and why?
4. Of a given list of technologies, which presents the greatest future opportunity to combat online child sexual exploitation and why?

Workstream 3: Literature Review

We began our literature review by defining the research question: identify safety by design interventions across each specific technology included in the ecosystem survey. To the best of our knowledge, this type of systematic overview of the available literature on this topic has not yet been conducted. We then compiled a list of sources (e.g., Google Scholar, arXiv.org, public documentation from child safety regulators) and a list of key terms to use (e.g., child safety, CSAM interventions) paired with key terms relevant to the particular technology (e.g., end-to-end encryption, virtual environments) to use in searching through these resources for relevant literature. Upon compiling an initial set of literature through the above process, we reduced this initial set to those that were most highly cited or otherwise came from known experts in this space. We then read and analyzed the literature, looking for patterns, trends, and gaps in the research.

Fig 1 | **Sector Breakdown**

Sector selection was multi-select, so the total may not match the total participant count.



Featured Technologies

As technology has developed and become embedded in our daily lives through hardware, such as smartphones, or applications, such as messaging or live streaming tools, digital tools and environments have opened new vectors of risk for children and new opportunities for those wishing to sexually exploit or harm them.

Through this research, the stakeholders and resources with which we engaged painted a complex picture: a myriad of technologies plays diverse roles in preventing and combating technology-facilitated child sexual abuse. These technologies rarely occur in complete isolation from one another and their impacts depend heavily on how and why someone uses them

The technologies selected for this report were those assessed to present the most significant potential for near- or far-term impacts to the issue of child safety. This includes technologies that hold significant promise to accelerate the work of investigators and content moderators, as well as those that risk unlocking new environmental risks and tactical advantages for offenders. This report should not be considered exhaustive and should be considered as one step in a broader process of understanding the opportunities or challenges posed by such technologies.

Importantly, the scale and nature of these technologies will change with time in response to the evolution of product features, further technological innovation, user popularity, and social factors such as behavioral norms and regional policies. At the time of this research, those in the field are observing cases involving many of these technologies, demonstrating not just their theoretical but real misuse. However, they do not yet report them to be the dominant case type.

In this section:

Predictive Artificial Intelligence

Technology Overview

Predictive artificial intelligence (predictive AI) is an umbrella term for a range of machine learning (ML) technologies that can recognize patterns and make predictions without being explicitly programmed to do so. There are three high-level categories of ML: supervised, unsupervised, and reinforcement learning.

In supervised learning, the model is provided with a set of training data, where the data is labeled according to the target categories, concepts, or information. The model iterates over the training data and, in this way, learns to place different weights on relevant features to distinguish the categories from each other, ultimately producing a weighted function that takes a piece of data as an input and outputs a prediction.

Unsupervised learning similarly involves a model iterating over training data to learn the target output; however, unlike supervised learning, the training data is not labeled. As a result, with unsupervised learning the model iterates over the data to find patterns and similarities between relevant features to organize the data into discrete groups.

Unlike supervised or unsupervised learning, reinforcement learning doesn't require training data. Instead, the model learns the "policy" by exploring its environment in real time. With each action a model takes, it either receives a reward — if the choice it takes is desired — or a punishment, if not. In this way, with successive rewards and punishments, the model learns the "policy": the strategy of whatever task it is that the model is being trained to accomplish.

> **"Predictive AI is already being used across investigations into online child abuse, to identify patterns of luring or grooming, to identify victims and perpetrators, and more. It is essentially the only tool available to help process huge amounts of information online."**
>
> *— Sector: Other; Tenure: 1-3 years*

Predictive AI is used across a number of different industries,[5] including health care,[6] marketing[7]/advertising, financial services,[8] and many more. Recommender systems[9] are powered by predictive AI and are used across a wide range of industries. Across all of these applications, there is ongoing cross-sector effort to establish[10] and in some cases regulate[11] best practices for ethical development and use of these systems, across a number of relevant categories[12] (e.g., bias mitigation, privacy and consent, transparency, accountability).

Current Opportunities & Impacts

In addition to the everyday use cases outlined above, technologies leveraging the power of predictive AI models are currently some of the most powerful sets of tools for combating technology-facilitated child sexual exploitation. The sheer volume of data exchanged online each day and the velocity of interactions requires technology to augment the capabilities of human moderators. Predictive AI complements legacy content-moderation tools, like hashing and matching, to surface, prioritize, and triage violative content — such as threats of self-harm or child sexual abuse material — for moderator review.[13,14] Further, predictive AI can support efforts to tackle offline child sexual abuse, specifically for victim identification. Classifiers may identify novel child sexual abuse material that depicts a previously

unknown child in an active abuse situation. They also can be used to support other victim identification workflows, such as prioritizing seized data for review and extracting information regarding the location of a victim depicted in an abusive image or video.

In addition, use of predictive AI unlocks significant opportunity to analyze large datasets. Not only does it make it possible to increase the insights we unlock and the pace at which we arrive at them (through larger volumes of data and by identifying connections more rapidly than human review), but also it may limit the amount of exposure to harmful content (such as chat logs involving child sexual exploitation) within that dataset.

It is worth noting that in those scenarios described above, building predictive AI will involve content exposure, as building these models requires human labeling and review of training data. The holistic cost of exposure may, in the long run, be reduced, as the amount of training data needed to be labeled may be less than the amount of input data that the model is then used to analyze, but it should be recognized that the cost of this exposure will move from the person conducting the analysis to the person labeling the training data.[15]

> **"AI supports law enforcement officers in reviewing CSAM and automatically prioritizes the most urgent ones — reducing the exposure of officers working in this field."**
>
> *— Sector: Government; Tenure: 3-5 years*

Alongside the ways predictive AI may prove a useful tool in the fight against technology-facilitated child sexual exploitation, several concerns exist surrounding the technology and its potential risks. In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector* indicated that they were handling CSAM cases involving predictive AI, ranging from 0% to 10% of their CSAM cases (with an average of 0.48%). Across this project, two particular themes of risks were surfaced:

- Algorithms' role in elevating harmful content and connections.
- Insufficient human checks and balances.

Algorithms' role in elevating harmful content and connections. While recommender systems powered by predictive AI can offer benefits in some use cases, they have also been shown to suggest risky connections and harmful content.[16] Recommender algorithms have a demonstrated history of elevating accounts belonging to minors for discovery by users apparently interested in explicit minor imagery.[17] In some studies, researchers who encountered potential child exploitation accounts saw recommendations for similar accounts following exposure to the initial account.[18] In addition, several studies have highlighted the spiral of negative content that users, including minor users, can be exposed to based on behaviors surrounding certain types of content.[19]

*The base size of this group was small (n<25), and therefore these findings should be considered directional.

Insufficient human checks and balances. The role of a human in the loop is critical to the safe and beneficial use of predictive AI. In fact, recent polling found a large majority of US voters are concerned about the impacts of AI making decisions without adequate human oversight,[20] and similar sentiment was echoed in the 2023 Bletchley Declaration.[21] Predictive AI does not carry the same ability to recognize contextual nuance as human experts, and insufficient human checks and balances can lead to negative outcomes. Models that are not maintained will degrade in performance, and bias in models may lead to unintended outcomes and missed situational interpretations. Reporting has highlighted instances where content moderation AI fails to perform effectively for certain groups, regions, or languages.[22] In this way, the role of human intelligence and interpretation — as well as clear and reliable mechanisms for contesting and appealing decisions[23] — remain crucial.

Future Opportunities & Impacts

Predictive AI has already been incorporated into our society in a number of different capacities, with both positive and negative outcomes and increasing regulatory attention. Yet, for those on the front lines of combating technology-facilitated child sexual exploitation, the potential benefits have not yet been fully realized. This is not the result of lack of will or desire; however, there is currently a lack of data specific to this risk vector, and there is still insufficient investment in the development of specialized tooling for deployment on this topic.

For example, the performance of classifiers is reliant on the quality and representative nature of the training data and associated labels. For classifiers intended for use on abusive interactions, performance will be improved by training on pertinent data (such as CSAM or sextortion interactions) that has been labeled with the necessary specialized expertise and within secure environments (such as within law enforcement facilities). The ecosystem's ability to leverage this powerful technology in service of combating child sexual exploitation will be heavily dependent on the availability of these two items.

Concerningly, many companies have made significant reductions to personnel, particularly in sectors relating to trust and safety and AI ethics, which may be a worrying sign for their investment in maintaining and evolving AI-powered moderation tools.[24,25] While AI solutions can scale and accelerate some tasks, in the child safety realm in particular, overreliance on technology in content moderation pipelines in service of cost-cutting measures will result in worse moderation decisions, worse model performance, and missed signals of abusive and exploitative interactions. Pathways for continuous feedback, regular evaluation, and iteration for model maintenance are crucial to account for adversarial changes in offender behavior, as well as model drift and bias. This type of continuous iteration requires human oversight and will likely involve additional labeling burden. As noted above, the wellness cost of labeling this type of data can be high. Ethically navigating this reality will include ensuring the humans conducting the labeling have robust wellness resources, that any unexpected content exposure is minimized, and they have the freedom to exit the labeling session when needed.

> **"The use of predictive analytics to profile and target children could lead to privacy violations and exploitation. The potential for biased algorithms to perpetuate harmful stereotypes and make incorrect predictions about children's behavior adds another concern."**
>
> *— Sector: Non-profit; Tenure: 7+ years*

Predictive AI may also be used in efforts to estimate or verify age, as a mechanism to prevent access to age-limited content or prevent interactions between children and adults. The use of this type of technology for these purposes is not yet ubiquitous, and its emergence has been paired with efforts to transparently benchmark its performance.[26]

Safety by Design

Efforts in safe design of predictive AI technology and predictive AI systems in the context of technology-facilitated child sexual exploitation and abuse primarily focus on the recommender systems, both in terms of their use for recommending content (media, advertisements, other user posts, etc.) and recommending connections. Some of the suggested interventions are applicable in both settings — for example, user empowerment via feedback loops[27,28,29,30]; transparency in implementation, objectives, and parameters[31,32,33]; regular auditing[34,35]; and changing algorithmic objectives (e.g., to optimize for a specific human value, as opposed to optimizing for maximal user engagement).[36,37,38,39,40] For content recommender systems, interventions include alternative curation models (e.g., chronological listing of content) [41,42] and more rigorous content filtering before the recommendation is surfaced.[43,44,45] Across both types of use cases for recommender systems, some of these interventions have further become codified in law.[46,47]

It is worth noting that there is a rich ecosystem of literature on interventions and best practices in predictive AI more generally (across a number of relevant categories, including bias mitigation, privacy and consent, transparency, accountability, etc.) There is clear overlap with this broader literature, and preventing predictive AI–accelerated child sexual exploitation and abuse, that could be worth further exploration.

Generative Artificial Intelligence

> **"Generative AI can be, and is, used to generate CSAM — sometimes by other children with the likeness of specific people like classmates, etc. There is a very significant amount of effort being made by bad actors on circumventing the protections put in place around generative AI models."**
>
> *— Sector: Technology; Tenure: Less than 1 year*

Technology Overview

Generative artificial intelligence (generative AI) is an umbrella term for a range of ML technologies that can generate new content (images, videos, text, and audio) without being explicitly programmed to do so. Like predictive AI, it sits within the same three high-level categories of ML (supervised, unsupervised, and reinforcement learning). Generative AI consists of the same fundamental building blocks as predictive AI, but rather than outputting a prediction on a target category or variable, it benefits from advancements in large language models and diffusion models to instead output human-understandable media.

Over the last few years, generative AI has been increasingly adopted across many of the same industries as predictive AI.[48,49] Standards and regulation for the ethical development and use of generative AI systems are still emerging,[50] some of which build off of existing standards for predictive AI. Within the broader conversation of AI safety, there are primarily two lines of thought: one that focuses on existential risks and one that focuses on present-day harms.[51]

Current Opportunities & Impacts

Generative AI technologies have seen a rapid acceleration in capabilities and everyday adoption in the last several years. The technology exists not only in stand-alone applications built and/or supported by dedicated generative AI companies but also embedded into a wide range of tools, services, and workflows, including writing assistance,[52] customized playlists,[53] and recipe generation.[54] Generative AI technologies can accelerate work and unlock new ways to learn, explore, and create.

These same capabilities impact the risks of technology-facilitated child sexual exploitation. AI-generated (AIG) CSAM is a now well-documented abuse of this technology.[55] In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector* indicated that they were fielding CSAM cases involving generative AI, ranging from 0% to 30% of their CSAM cases (with an average of 6.66%).

Several categories of risks have been associated with this technology. These impacts can be summarized into a few specific areas of risk, as described in Thorn and All Tech Is Human's report *Safety by Design for Generative AI*[56]:

- Creates new ways to sexually exploit and revictimize children.
- Enables misuse of children's benign and abuse imagery in training models.
- Reduces social and technical barriers to sexualizing minors.
- Impedes victim identification.

*The base size of this group was small (n<25), and therefore these findings should be considered directional.

Creates new ways to sexually exploit and revictimize children. Generative AI models have been used to create CSAM depicting both new and historical victims. In cases involving historical victims (those whose original abuse imagery did not include the use of generative AI technologies), offenders have been reported to generate media that shows new forms and levels of sexual violence inflicted on a historical victim. They've also been reported to create imagery that inserts their own likeness into historical abuse events.[57]

Generative AI models have also been used to generate child sexual abuse imagery of new children. This has been seen in instances of offenders creating abusive imagery of children that may be known or unknown to them. In some cases, offenders have used benign imagery, acquired from public sources online or through direct access via family or friends,[58] or captured surreptitiously while encountering the child in a public place, as reported in the recent investigation of Justin Culmo.[59]

These cases are also not strictly limited to those interested in sexually abusing a minor. Financial sextortion — attempts to extort a victim for money by threatening to leak nude images — has found extorters using generative AI technologies to scale their efforts.[60,61] In addition, several cases have been reported involving use of generative AI apps by minors to create "deepfake nudes" of other kids.[62] This is discussed further below.

Enables misuse of children's benign and abuse imagery in training models. Prominent models designed for mainstream use cases have been identified as the tools used to produce AIG-CSAM. Models are at greater risk for producing photo-realistic AIG-CSAM when their training set includes both sexually explicit imagery of adults and benign imagery of minors. In this way, at the direction of the user, the model may be able to combine these concepts, based on the contents of its training set, to output sexually explicit images of minors.[63,64] In addition, the training set involved in the most often recognized model involved in generating AIG-CSAM (Stable Diffusion 1.5) was determined to include confirmed CSAM.[65] In both instances, children are directly involved, both via historical abuse imagery and everyday online photos, in the creation of additional abuse imagery of children.

> **"Adolescent perpetration of CSA of other children is on the rise. There are currently 30-some phone apps that are easily accessible to generate AI-images of nudes or CSA using children as subjects. The combination of both these problems exacerbate instance of CSA. Even if not actual bodies or CSA, they can affect great trauma on chidren who are subjected to this. In addition, the facility/access to this technology and the normalisation of sexual objectification is similar to the normalisation of violent porn and influences young people's attitudes about sex, consent, and sexual violence."**
>
> *— Sector: Non-profit; Tenure: 3-5 years*

Reduces social and technical barriers to sexualizing minors. While cases of offenders using photo-editing software to manipulate existing images to create new abuse scenes are not new,[66] the output, in the absence of significant time and ability, was often crude and obviously manipulated. Generative AI reduces the technical barriers to producing custom abuse imagery of minors rapidly and easily, putting the tools into the hands of anyone with an interest. The ability for users to not only create imagery easily but to produce material that mirrors their unique abuse fantasies in the acts and victims targeted may serve to reinforce those thoughts and feelings.[67]

Generative AI technology not only makes it easier for those intent on sexually abusing minors to create custom material, but abuse of the technology is also occurring among minor peer groups to create bespoke deepfake nudes of classmates. In a recent survey of youth aged 9-17, 1 in 10 minors reported their peers had created deepfake nudes of other peers.[68] Similar cases have been reported across a number of schools, including in New Jersey,[69] Spain,[70] and California.[71] While the motivation in many of these cases likely differs from what motivates adults with a specific sexual interest in minors — rather, it is likely often the result of poor judgment or awareness of victim impact by minors — the outcome is still a form of image-based abuse.

Currently, sentiment is mixed in some communities regarding the harm of this behavior and the best interventions, with public response ranging from minor prosecutions[72] to viewing them as "youthful transgressions that should be forgiven."[73] This creates multiple challenges. On the one hand, criminalization without appropriate investment in prevention or restorative interventions may lead to more harm than good.[74,75] On the other, equivocation about the permissibility and harm caused by creating nude images of someone without their consent fails to equip young people with the awareness necessary to avoid causing harm via this novel technology and fails to honor and support victims experiencing this abuse.

Impedes victim identification. The systems responsible for the moderation, triage, and investigation of events involving child sexual abuse are already notably strained.[76] Much as generative AI technologies can support the rapid production of synthetic media more generally, so too is the case for AIG-CSAM. The systems in place for responding to this content risk seeing a critical increase in volume, delaying vital services to impacted victims.[77] Beyond the volume itself, in cases where generative AI technology is involved, the process for disentangling such abuse imagery from that of children in active abuse situations (e.g., where the abuser/producer of abuse imagery has direct and present contact with the victim) becomes increasingly complicated. Investigators must assess what elements in an image (or its entirety) are generated or nongenerated. Challenges involving the manipulation of images by offenders to obscure critical identifiable details are not new[78]; however, the introduction of generative AI technologies makes it easier and faster to accomplish their goals.

> **"As I have mention earlier AI can be used for education and learning, parents can utilize these software to learn how to protect their children online."**
>
> *— Sector: Government; Tenure: 1-3 years*

Future Opportunities & Impacts

As was noted in the predictive AI section, the application of this technology in a child safety professional application (in other words, nonprofits, trust and safety, law enforcement, and others) lags behind general public usage. The use of off-the-shelf models for general learning and exploration of topics relating to child sexual abuse are (rightly) limited. For example, when instructing certain large language models (LLMs) to respond to specific questions about published academic articles relating to technology-facilitated child sexual exploitation, on several occasions, the requests were denied based on apparent conflicts with model policies. Prohibition on user queries concerning offender tactics is a common and important guardrail for models to reduce the ease with which someone looking to harm minors can quickly surface tips and strategies for abusing children online. However, those child safety professionals looking to apply the benefits of generative AI in their workstreams will, therefore, require more customized solutions.

In this vein, some organizations are exploring the safety, value, and viability of leveraging such technology to improve offense deterrence resources.

The Lucy Faithfull Foundation has reported promising outcomes from its chatbot to deter and redirect individual's search for child sexual abuse material online.[79] In its early version, the chatbot is limited to scripted responses, but the organization is exploring opportunities to update the model using LLMs to deliver more specialized and responsive deterrence messaging.[80]

For both of the above use cases, known issues with generative AI regarding model hallucinations will need to be addressed to ensure both the reliable retrieval of information and the generation of consistently relevant and appropriate text. These technologies are fundamentally statistically based, which means that out-of-the-box solutions treating them like a database will not be sufficiently rigorous, and reliable outputs for conversational purposes are not guaranteed.[81,82]

Generative AI technologies are evolving rapidly, and several additional risks have been identified based on the current trajectory of this technology and its potential for misuse. Both video and audio creation have seen significant improvement in quality over the last 2 years. Both of these formats are anticipated to be at risk of misuse for the exploitation of minors in the near future. And while thus far, AIG-CSAM remains but a small percentage of the overall CSAM in circulation,*[83] the ease and velocity with which it can be produced means concerns of overwhelming pipelines remain high.

In addition, generative AI technology poses an increasing risk to certain visual identity assurance techniques, such as checking a recorded or live image of the user. Within fraud monitoring circles, reports flag the capabilities of generative AI being at, or on track, to create significant disruptions.[84] These same tactics can undermine online child safety (e.g., creating challenges for certain age-assurance mechanisms used to enforce platform age gating).

Lastly, the parallel development of wearable technologies intersecting with the already observed misuses of generative AI for the purposes of abusing or exploiting children opens new avenues for risk. The ability to discreetly record children in public reduces existing protections against misuse of their images. Although features in current wearables attempt to alert to filming activity,[85] simple hacks may circumvent systems likely designed to balance aesthetics with function and may not sufficiently prioritize safety. This, intersecting with the capabilities afforded offenders by the abuse of generative AI models, opens increased opportunities for abuse imagery of children to be produced.

Safety by Design

A significant amount of research and exploration into building and deploying safe generative AI systems and models, across a variety of current and future harms, is actively underway. When focusing on safe generative AI specifically through the lens of child sexual exploitation and abuse, a broad range of interventions have also been proposed and explored at various stages in the life cycle of the model and across various actors in the AI value chain.

At the model level, these interventions include training data cleaning and curation (e.g., removing CSAM from training datasets),[86,87,88,89,90] children's data protection and privacy,[91,92] red teaming for technology-facilitated child sexual exploitation,[93,94,95,96,97] content provenance and disclosure mechanisms,[98,99,100,101,102,103] model capabilities testing and evaluation,[104,105,106,107] user reporting and feedback,[108,109,110,111,112] and transparency (e.g., on model capabilities and limitations,[113,114] as well as for datasets[115,116]). Of these interventions, requirements for content provenance and disclosure mechanisms have been codified into law via the European Union Artificial Intelligence Act.[117,118]

*Recent reporting states that the National Center for Missing & Exploited Children receives roughly 450 reports regarding generative AI, out of the total 30 million reports, monthly to their tipline, or .0015% (see endnote 49).

When considering deployment settings, interventions such as closed-source model input and output content moderation,[119,120,121,122,123,124] harmful model removal (e.g., de-indexing CSAM fine-tuned models from search results),[125,126] synthetic content detection and AIG-CSAM hashlists,[127,128] and age-appropriate design[129,130] have all been suggested.

Much work remains to iterate, refine, assess, and maintain the efficacy of these interventions. In the meantime, the ecosystem of generative AI models and associated tools continues to grow rapidly.

End-to-End Encryption

Technology Overview

End-to-end encryption (E2EE) is a method of securing data such that only approved users can access the content of the data. E2EE can make use of both asymmetric cryptography (where two different keys are used to encrypt and decrypt the data) and symmetric cryptography (where the same key is used to encrypt and decrypt the data). E2EE ensures that even if the data is intercepted during transmission, it cannot be decrypted and read by third parties (as long as those third parties do not have access to the cryptographic keys). E2EE is most commonly used to secure communications between users: encrypting the relevant data before transmission from the sender and decrypting the data only after receipt by the intended recipient. In that use case, after receipt and decryption of the data, the data is further verified to confirm its authenticity (i.e., confirming the message was sent by the alleged sender by verifying their digital signature) and its integrity (i.e., confirming the data was not tampered with during transmission).

As noted above, E2EE is most commonly used to secure communications (e.g., Signal, WhatsApp, ProtonApp) in order to protect user data and privacy. Other use cases include secure data storage (e.g., iCloud[131]), password management (e.g., 1password[132]), and securing financial transactions (e.g., Stripe[133]). While E2EE is implemented in various forms across different platforms, its use and adoption are not yet ubiquitous.

Current Opportunities & Impacts

E2EE offers valuable protections for individual data, regardless of whether the user is a journalist, dissident, or any individual concerned about the privacy and security of their communications, financial transactions, and personal images. However, these tools have also been weaponized by offenders looking to traffic in child sexual abuse material, groom children, and isolate victims from the protective features built into public spaces online and off.

> **"End-to-End Encryption (E2EE) remains essential for safeguarding the privacy and security of children's online interactions."**
>
> *— Sector: Non-profit; Tenure: 7+ years*

In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector* indicated that they were fielding CSAM cases involving E2EE, ranging from 0% to 80% of their CSAM cases (with an average of 26.19%). Across this research, several categories of risks surfaced. Core ways E2EE is impacting technology-facilitated child sexual exploitation include:

- Limited opportunities to deploy technology-facilitated child sexual exploitation detection tooling at scale.
- Heavy reliance on user reporting increases burden on victims of on-platform abuse.
- Criminals weaponize privacy to hide the trade of abuse material and isolate victims.

Limited opportunities to deploy technology-facilitated child sexual exploitation detection tooling at scale. Currently, hashing and matching is the dominant tactic to detect and remove child sexual abuse material at scale. In 2023, the National Center for Missing & Exploited Children received more than 36 million reports relating to online child sexual exploitation and abuse, including more than 100 million image and video files.[134] Electronic

service providers make up the lion's share of CSAM reports, having reported 35.9 million reports in the same period of time.[135] Hashing and matching plays a critical role in detecting these tens of millions of files and reducing the ongoing victimization of the children whose abuse images are being recirculated.

> **"E2EE messaging apps are utilized by offenders to communicate with minors, enticing them to produce sexually explicit material and going undetected by platforms. We regularly receive reports where the "meet up" takes place between a minor an adult on a non-encrypted platform and the conversation quickly moves to an encrypted channel."**
>
> *— Sector: Non-profit; Tenure: 7+ years*

In addition to hashing and matching, advances in applications of predictive AI (discussed earlier in this report) have equipped platforms with the ability to identify high-risk interactions and media indicative of child sexual exploitation and abuse. Current models enable platforms to detect high-risk CSAM and abusive interactions at scale, specifically trained to detect interactions involving sextortion, exchanges of CSAM, or likely offline child sexual abuse.[136,137]

It is not currently technically feasible to deploy hashing and matching or predictive AI solutions in fully E2EE environments, given today's available protocols and content moderation technologies, without compromising user expectations of privacy. This removes a vital technological pillar for combating technology-facilitated child sexual exploitation at scale. As framed in a human rights assessment conducted by Business for Social Responsibility, E2EE "does not inherently limit children's right to safety, but does make detection harder."[138]

Heavy reliance on user reporting increases the burden on victims of on-platform abuse. It follows then, and this is supported by commentary from platforms,[139] that user reporting will be a heavily-relied-upon mechanism to alert a platform of abusive (and potentially criminal) behavior on their services.[140]

The inherent flaw as it relates to technology-facilitated child sexual exploitation must be acknowledged. Most users witnessing CSAM or child sexual exploitation occurring in an E2EE environment will fall into one of two groups: (1) a child being victimized or (2) a user participating in the trade of CSAM or exploitation of minors. It would be against the interests of the latter to report, placing a significant burden on a child being victimized to seek help.

Importantly, many children do not disclose they are experiencing sexual abuse. For those who do, it may be far later in life and/or to friends rather than a trusted adult or authorities.[141,142]

> **"Using 'E2EE' as an excuse to stop detection makes it impossible to combat the latter."**
>
> *— Sector: Non-profit; Tenure: 3-5 years*

When a minor experiences a potentially risky sexual interaction online and chooses to take action, they are more likely to use online safety tools over disclosing to someone in their offline lives, such as caregivers or peers.[143] However, online safety tools like reporting and blocking are not always intuitive or effective. Reporting on ban evasion (where a deplatformed user attempts to regain access via a new account or other tactics) has highlighted the inconsistent success of keeping banned users off a platform.[144,145]

In addition to the limitations of reporting, minor users indicate they are more likely to simply block an abusive account rather than report it to the platform. Research highlights that these interventions are viewed differently by minors, where blocking is viewed as a mechanism to stop contact, and reporting is viewed as a tool to get someone "in trouble".[146]

User reporting alone, for the reasons stated above, is unlikely to surface the same volume of alerts for child sexual abuse and exploitation occurring in E2EE environments as tools like hashing and matching or predictive AI classifiers. However, for those who do use this feature, significant work is needed to ensure the reporting process is clear, accessible, and effective.

Criminals weaponize privacy to hide the trade of abuse material and isolate victims. Offenders often target environments based on risk-reward calculations. An environment where they have a reduced likelihood of detection or being reported while engaging in criminal activities — such as grooming child victims or exchanging child sexual abuse material — holds obvious value for them.

In offline contexts, this often shows up with child sexual offenders pursuing positions of trust within the community or family to gain access to children and their adult support networks.[147] They pursue impressions of trustworthiness within public settings while moving the child to more isolated situations to avoid observation by others. In this situation, attempts to disclose abuse by the victim can be met with disbelief, placing the burden on the child victim to somehow prove the offender is not who they appear to be in public ecosystem.

Similar practices have been observed within online spaces — wherein offenders leverage more public services, such as gaming and social media sites, to find and initially connect with potential victims before moving them to private messaging services.[148] In a survey among American minors aged 9-17, two in three reported having been asked to move from a public forum to a private chat by someone they had met online.[149] This behavior has also been reported in financial sextortion cases. A recent analysis of apparent financial sextortion reports made to the CyberTipline found extorters moving victims to secondary platforms. While many more social platforms play a dominant role in where initial contact is made, messaging services, including those with perceived increased security via E2EE such as WhatsApp, iMessage, and Telegram, featured far higher in the list of platforms the victim was moved over to for continued interaction.[150]

Future Opportunities & Impacts

E2EE promises to deliver important privacy protections to many as its adoption increases. Specific to online child safety, some survey respondents elevated its value in protecting children's data, including their personal images and financial information.

And while this may be true, we also know that platforms which have played a significant role historically in proactively detecting child sexual exploitation and abuse have either recently, or are actively, implementing encryption across their systems, a move that may impact the volume and/or quality of reports made. Meta's Messenger service is perhaps one of the most publicly discussed of these cases. When Meta announced in December of 2023 their intention to move Messenger to full E2EE,[151] those in the child safety ecosystem expressed considerable concern that this move will remove critical safeguards in an area known to house a high number of reports concerning technology-facilitated child sexual exploitation.[152] In 2023, Meta was responsible for more than 17 million reports to the

*Meta is listed as "Facebook" in this source.

CyberTipline (nearly half of all reports made by electronic service providers to the National Center for Missing & Exploited Children during this time).[153]* It is too soon to tell how much the move to encrypt this service will impact the overall volume of reports or the quality of the report contents.

Safety by Design

Given that E2EE is specifically designed to secure user data, it is not currently technically possible for a service provider to detect, remove, and report CSAM in fully end-to-end encrypted environments with existing E2EE protocols and content moderation capabilities without compromising user expectations of privacy. However, given the significant risk of abuse of encrypted environments for child exploitation, it is critical to explore platform design strategies that maintain the security of encrypted information while reducing vulnerabilities to this abuse. Such strategies may focus on steps that can exist alongside encrypted exchanges rather than within the encrypted data itself, or may be aligned more closely with platform policy decisions than E2EE specific design strategies.[154]

Much of the existing literature in this space focuses on the various tradeoffs at hand when attempting to detect CSAM in end-to-end encrypted systems. Recommendations and guidance tend to come paired with caveats and limitations, with some literature focusing on highlighting the limitations of particular strategies.[154]

For interventions related to CSAM detection within and around end-to-end encrypted systems, interventions as a whole can be characterized in two ways: those that do not rely on automated detection and those that seek to make use of automated detection in a way that preserves user privacy. For the former, suggested solutions include user reporting,[155,156,157,158] age verification,[159] limiting the capacity for content sharing,[160,161] and user education and prompting.[162,163,164] Some literature further explores verifiable user reporting (e.g., message franking).[165,166] For interventions related to privacy-preserving automated detection, hashset transparency and auditability,[167,168,169] metadata analysis,[170,171,172] privacy-preserving detection (e.g., homomorphic encryption,[173] private membership computation,[174] private set intersection[175]), and client-side scanning,[176,177,178] have all been suggested and debated. These debates have introduced questions about changes to the underlying technology and how those changes impact the inherent positive benefits of E2EE.

*The base size of this group was small (n<25), and therefore these findings should be considered directional

Extended Reality Environments

> **"I can see XR enhancing some of our online training modules to make them more engaging and experimental"**
>
> *— Sector: Non-profit; Tenure: 7+ years*

Technology Overview

Extended reality (XR) is an umbrella term that encompasses virtual reality (VR), augmented reality (AR), and mixed reality (MR) technologies. All three of these share the fusing, or blurring, of lines between the physical and virtual worlds. In VR, that occurs via complete immersion into a digital environment. In AR, physical reality is enhanced and/or overlaid with elements from the digital worlds. In MR, elements of both VR and AR are combined such that users can interact with both physical and digital objects at the same time.

XR requires a host of technologies, in particular specialized hardware, such as motion sensors (to track a person's head, body, or hand positions), gyroscopes (to maintain a user's orientation), small screens (for 3D display), other sensory peripherals (e.g., controllers, headphones), etc. Software for spatial computing and mapping and 3D modeling and rendering are also used to enable the creation of realistic environments and power interaction with the physical world within MR and AR environments. At a high level, there are two main categories of XR devices: stand-alone — where all the components necessary for the experience are in the headset (such as Oculus VR) — and tethered — where the headset is connected to some other device like a video game console (such as SteamVR).

The adoption of XR is still in its early stages, with projections of continued growth. It has applications in a number of different industries, including gaming,[179] education,[180] advertising,[181] and manufacturing.[182]

Current Opportunities & Impacts

XR experiences are appearing across a number of sectors, including gaming, retail, travel, education, and even adult entertainment. The technology may increase engagement,[183] unlock new geographic and intellectual opportunities,[184] and offer new pathways and tactics for counseling and therapy.[185] For those facing physical or psychological barriers to travel, socialization, or otherwise interacting outside of the home, XR technologies stand to be transformative.[186] The adult content market in has shown notable growth in recent years, with forecasters predicting continued opportunities within the VR sector.[187,188]

> **"...XR is an unregulated space and we still do not understand the impact of virtual abuse on the recovery of children..."**
>
> *— Sector: Non-profit; Tenure: 7+ years*

Although adoption of XR remains in its early stages, cases have been reported[189,190] and studies have been conducted indicating the ongoing presence of XR-facilitated child sexual exploitation.[191] Some of these reports demonstrate the potential interplay between offline and online abuse, where the immersive nature of these technologies can be used to distract observers from physical child sexual abuse that is actively occurring.[192] These reports may grow alongside a forecast of increasing minor users in response to reduced headset costs and lowering of recommended user ages.[193]

In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector indicated that they were fielding cases involving XR, ranging from 0% to 10% of their caseload, with an average of 0.95% of their caseload.

During this research, three categories of risks emerged. Harms already observed include:

- Limited practices to reduce exposure of minors to adult scenarios and interactions.
- Increased opportunity for exchanges outside of dominant content moderation tooling.
- Immersive nature of the environment can disrupt users' understanding of risk and harm.

Limited practices to reduce exposure of minors to adult scenarios and interactions. XR environments have made progress in implementing parental control capabilities to defend minor users from encountering adult content. Reporting in 2022 raised concerns about the lack of ability to block 18+ content.[194] Some devices have since added such controls, creating more ways for parents to curate the types of content minor users can access and the process through which they can approve new contacts.[195]

Unfortunately, as with other app stores, content maturity ratings are assigned by the developers and may not match the user experience. For example, VRChat is listed for users 13+; however, several reports show minor encounters with adult content,[196] such as was reported by a BBC reporter who, while posing as a 13-year-old user, was able to visit a virtual strip club in the app.[197] Similar experiences have been reported in other environments, such as RecRoom and Horizon Worlds.[198]

Increased opportunity for exchanges outside of dominant content moderation tooling. Content moderation in traditional 2D environments carries its own challenges. However, expanding to immersive environments introduces additional policy and technical hurdles. While traditional content moderation tactics, such as hashing and matching, and classifier review can support review efforts for some media in XR environments, many interactions exist outside of traditional image/video upload or text-based exchange.

Current content moderation capabilities for ephemeral events, such as audio or embodied interactions, are limited compared with the body of moderation tools designed for logged interactions, such as uploaded images, videos, or text. In addition to the lack of technology to identify violative behavior in real time, the lack of a record of the interaction limits the verification of a reported infraction or enforcement of a policy in some instances.[199,200] As a result, the potential for interactions outside of the protections of existing content moderation capabilities is much higher in XR environments than in traditional 2D social media exchanges.

Immersive nature of the environment can disrupt users' understanding of risk and harm. Extended reality environments are designed to blur the user's awareness and boundaries between what physically surrounds them in their offline space and what they observe and experience via the virtual environment they're operating in. This feature enhances the user's experience, helping them to more completely believe they are in the virtual environment. However, this can also lead to less clarity about the emotional and psychological impacts of experiences had in these virtual worlds.[201]

Reporting has highlighted the tension that exists between a virtual world and the "realness" of the experiences had within that world. In some cases, these experiences of virtual groping, harassment and rape occur within minutes or an hour of entering the virtual environment.[202,203,204]

> **"…lack of awareness regarding the impact of harmful behaviours within the VR/XR/AR space. It's not at the high end of the scale spectrum yet but research and lack of guardrails heeds caution at the proximity between trauma in these mediums vs in real life. The range of harmful behaviours ranges from hurtful language/ bullying right through to virtual sexual assault so the spectrum for harm is wide."**
>
> *— Sector: Other; Tenure: 1-3 years*

In one example, gamer Jordan Belamire shared her experience of being virtually groped and harassed. She described being followed and repeatedly groped despite telling the harassing user to stop. "This goaded him on, and even when I turned away from him, he chased me around, making grabbing and pinching motions near my chest. Emboldened, he even shoved his hand toward my virtual crotch and began rubbing."[205]

VR users have reported mixed opinions regarding how they feel about the experiences had in virtual environments, where some report a clear distinction between fantasy and reality, and others report the nature of the interaction creates very real emotional and psychological responses.[206] In an incident first reported in 2023 and currently under investigation, a minor was online in a VR environment when several male players attacked and "gang-raped" her digital avatar.[207] According to the *Daily Mail*, a senior official familiar with the case noted that "this child experienced psychological trauma similar to that of someone who has been physically raped. There is an emotional and psychological impact on the victim that is longer term than any physical injuries."[208]

There are ongoing debates within the public regarding whether this type of abusive behavior is harmful or should be acceptable.[209,210] In one analysis of Reddit threads discussing the experiences Belamire described, opinions were split, with one Redditor commenting, "My political correctness is being pushed beyond its limits here. I'm sorry but there is no sexual harassment in multiplayer games. There are horny teenagers that get excited they have the courage to interact with a female." And another, "Of course today's VR experiences are a pale reflection of reality — but in no time at all, they'll be so realistic, you won't be able to brush it off."[211]

Notably, the comments studied in the above report were made between 2016 and 2017. Advances in XR technologies have only increased the "realness" of experiences, further pressing the issue of the psychological and emotional impact of XR experiences.[212] With XR pornography growing, it is likely that haptics will continue to be incorporated into these experiences,[213] offering risk of intersection with child sexual exploitation and abuse.

Future Opportunities & Impacts

To date, the adoption of XR technologies is still in the early stages, likely at least in part due to the high cost of devices.[214] However, as device cost drops, adoption may increase, along with interest to invest in further development. This cycle of adoption to new development, to new users, to further adoption will also carry opening for novel risk. Although there are still several unknowns (including the scale of adoption around minor users), the combination of factors described previously appears to make an environment of this nature particularly high risk.

Safety by Design

Efforts for safe XR design with children in mind cover a broad range of goals, from child safety, to child security, to privacy of children's data. While a significant portion of XR safety research is focused on adults, some of

the recommendations are age agnostic and therefore potentially applicable to children. Across the multiple goals of security, safety, and privacy, the corresponding approaches are sometimes overlapping. They can be broadly categorized as feature-level interventions under three primary categories: empowering children to quickly restore safety in unsafe XR situations, ensuring age-appropriate experiences for children, and preventing the unnecessary dissemination of information and data about a child or their environment. Examples of interventions in the first category include easy functionality (e.g., quick reactions) for blocking and reporting,[215,216,217,218,219,220] exiting a virtual space,[221] and personal boundary setting (e.g., proximity-based audio muting).[222,223] In the second category, recommendations include identity and/or age verification,[224,225,226,227,228] establishing spaces just for children,[229,230] time limits,[231] age-appropriate education and onboarding,[232,233,234,235] and support for parents and caregivers.[236,237,238,239] Data privacy and security recommendations cover a broad spectrum, inclusive of general best practices, like transparent and minimal data collection practices,[240,241,242,243] and specific interventions, like enabling privacy in shared spaces (e.g., via profile visibility control).[244]

Where current research seems to be lacking is an exploration in automated content moderation for detecting child sexual abuse and exploitation in these spaces, with some resources highlighting the challenges of automated content moderation in XR.[245,246]

Decentralization

Technology Overview

Decentralized computing refers to an approach to networking where computing tasks are split across multiple different entities, rather than one centralized authority. Decentralized networks make use of communication and message protocols to route traffic and interface between nodes in the network. Depending on the application, these nodes can be computers, servers, or other devices. The nodes in a decentralized network are able to directly connect with each other in a peer-to-peer (P2P) capacity. The tasks conducted within the network (whether that be communications, data transfer, or some other computation) are not coordinated by a central authority but rather established independently by the nodes. There are a variety of protocols that can be used for decentralized networks depending on the application, such as the P2P protocol (enables nodes in a network to request and respond to data) and ActivityPub (enables the servers in a decentralized social network to deliver content and other user activities to each other).

Decentralized computing can be applied across a wide range of applications, including decentralized social networks,[247] data storage,[248] financial transactions,[249] and machine learning.[250] Across all of these applications, decentralized networks allow for better fault tolerance and increased privacy. While decentralized networking is implemented in various forms across different applications, its use and adoption is not yet ubiquitous. Depending on the application, there are multiple challenges to designing and adopting decentralized networks, including resource scheduling, global policy enforcement, and reliable data availability.

Current Opportunities & Impacts

Decentralized networks have been in use for decades, but have started to see more mainstream attention in the last 10 years tied to the rising popularity of bitcoin and, more recently, the fediverse (a "decentralized group of social media platforms in which each independent platform can interact freely with any other platform that is part of the group"[251]). In distinct use cases of this core technology, both of these have a record of intersecting with risks of technology-facilitated child sexual exploitation and abuse.

In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector* indicated that they were fielding CSAM cases involving decentralization technology, ranging from 0% to 50% of their CSAM cases (with an average of 4.29%). During this research, several categories of risks were surfaced. The main areas of observed risk to date around decentralized networks include:

- Increased investigative challenges for suspect identity.
- Lack of centralized authority for standards and moderation.
- Difficulty removing illegal material.

Increased investigative challenges for suspect identity. One of the core features of decentralized networks is the enhanced privacy for users. Transactions do not involve a centralized authority (such as a financial institution), and accounts often do not require a real-world identity. Cryptocurrencies, one application of decentralized networks, have been associated with a variety of crimes in recent years, including CSAM, and per a report from the Internet Watch Foundation, their use in payment for

*The base size of this group was small (n<25), and therefore these findings should be considered directional.

purchasing CSAM has been on the rise.[252] While cryptocurrencies were at one point broadly perceived as untraceable, investigators have successfully identified offenders behind cryptocurrency-related CSAM offenses.[253] However, analysis of the use of cryptocurrency by CSAM sellers points to the likely appeal of privacy-centered coins in particular, such as Monero, in attempts to shore up privacy.[254]

Lack of centralized authority for standards and moderation. While removing a centralized authority supports the desired increased autonomy and content freedom among users of decentralized networks, this also removes a critical element of traditional trust and safety frameworks that have developed through mainstream centralized networking systems. Centralized authorities traditionally drive policies about acceptable use (and, by extension, prohibited use). Many platforms (in response to both social commitment and legal requirements) also deploy and maintain tooling necessary to enforce these policies at scale. In the absence of this centralized authority, abusive behavior and illegal transactions (including the trade of child sexual abuse material) can prosper.[255]

> **"Decentralization could make it easier for perpetrators to hide online material and activities that put child safety at risk — it is an opportunity to bypass current protections."**
>
> *— Sector: Other; Tenure: 5-7 years*

Lacking a centralized network does not, however, mean the operators and users of decentralized networks accept seeing hostile or illegal activity flourish within their decentralized communities. Rather, a variety of alternatives arise — oftentimes driven by volunteers committed to the health of their community. Unfortunately, content moderation work, as is gaining increasing public attention, can lead to significant exposure to violent and abusive material, including child sexual abuse material.[256,257] Research and reporting have documented the serious outcomes experienced by many exposed to this material, including anxiety and depression.[258]

Larger companies have encountered public pressure in the last 10 years as awareness of the impact of this work on moderators has increased. Many have implemented additional safeguards, including content moderation tools to help triage the massive volume of content, as well as wellness-enhancing strategies, such as obscuring elements of the image where visual review is not necessary to make a determination, or more generally increasing investment in wellness resources for employees performing this work.[259]

However, those volunteers performing similar work for decentralized networks often have none of the training, tooling, and support that is financed in large, centralized companies, leaving them ill-equipped to manage the significant emotional and psychological toll of exposure to content relating to child sexual exploitation and abuse.

Difficulty removing illegal material. One of the draws for decentralized networks is the security and reliability that stems from information secured in the blockchain. Decentralized recordkeeping makes tampering with — changing or removing — records much more difficult. This same quality applies to illegal material that has either been federated or recorded to the blockchain. While authors are unaware of any confirmed instances of CSAM saved to the blockchain, research has recorded the challenges of successfully removing CSAM across federated networks. As discussed by one of the authors of the Stanford Internet Observatory report on CSAM in the Fediverse,[260] remote instances of the detected CSAM in some cases remained beyond initial reporting to local instance operators. This may be tied to the independent role of operators in sending "delete events" to alert

operators or it could be a reflection of slow (or no) action on behalf of some remote operators where the content was not removed. While the offending post or user can be removed by an instance operator, the abusive or illegal content may be federated far beyond the initial instance, resulting in a spider web of impacted servers.[261]

Future Opportunities & Impacts

In looking to the future, most concern relies not around novel areas of risk, but rather increased adoption of decentralization without adequate checks and balances in place for the already observed risks. Researching any form however, some novel opportunities may also lay ahead that could benefit those working to combat technology-facilitated child sexual and abuse.

Researching any form of online criminal activity or abusive behavior has carried challenges; however, social media has played an increasing role in understanding network behaviors and abusive trends. Studies on hashtag trends have highlighted both the prevalence of some underrecognized harms and the gaps in platform efforts to block such activity.[262] Studies on public forums have elevated the experiences of survivors of sextortion.[263] Research on decentralized networks, such as Mastadon, is far from new. However, few studies have looked at the issue of child sexual exploitation and abuse. While decentralization does not mean all information is publicly discoverable (and many identified as open to child sexual exploitation may either be hidden or widely defederated[264]), this may be an area of increasing opportunity to examine network behaviors and public discourse on topics that may be more heavily moderated on other centralized platforms. It is worth noting that any such findings may not be generalizable to the broader public, as choosing to make use of a decentralized network may indicate perspectives, biases, and influences that differ from those present in other segments of the public.

> **"A more decentralized network means more robust communications and avoids single points of failure, this is not only benefiting children but all of society."**
>
> *— Sector: Technology; Tenure: 3-5 years*

While the decentralized nature of these networks means policies and associated enforcement practices will differ greatly across networks, the grassroots nature of trust and safety could also lend to increased trust in the function of trust and safety. Child safety experts recognize the critical role of trust and safety teams. However, these teams have, at times, been dismissed as unnecessary cost centers on one end, and at the other, tools for platform overreach. These types of characterizations underestimate and understate the critical role of trust and safety teams in ensuring online environments are safe and healthy. Particularly through the lens of technology-facilitated child sexual exploitation, where the scale and nature of these abuses on mainstream platforms can be minimized, policies implemented to combat them at scale have, at times, been framed as heavy handed and even a form of censorship.[265,266] By situating content moderation with specific instances and outside of a central authority, there may be greater user buy in and transparency about the necessity for certain policies (such as those associated with technology-facilitated child sexual exploitation and abuse) and the tactics for enforcement of those policies.

Safety by Design

Current efforts around the safe design of decentralized technology as it relates to child sexual exploitation and abuse primarily focus on two settings: decentralized media storage and retrieval (e.g., peer-to-peer file sharing), and decentralized social media. In the latter, existing literature offers a range of possible solutions to enable content moderation for CSAM. Community moderation is the assumed path forward, so suggested

solutions tend to focus on pathways to support that existing strategy. Such solutions include server-level and community-level blocklists used by individual community nodes,[267,268,269] pluggable and shareable automated detection technology,[270,271,272] and media removal.[273] Other researchers suggest instance- or gateway-level moderation.[274,275]

For decentralized media storage and retrieval, solutions tend to focus on strategies to support or prioritize law enforcement investigations — such as file tagging,[276] targeted peer removal,[277] or use of detection technology.[278] There is a line of research around preventing the distribution of copyrighted files, that while not directly focused on CSAM prevention, includes strategies that may be similarly applicable.[279]

Where current research seems to be lacking is an exploration in making these spaces inhospitable for sexual harms against children to begin with, with relevant work focusing on the possibility of shifting norms by rewarding positive user behavior.[280,281] By and large, most research is focused on enabling detection, moderation, and prioritization in spaces where the abusive content and behavior is already occurring.

Quantum Computing

> **"Quantum Computing: stronger security through unbreakable encryption to protect children's data and future-proof online safety against quantum threats."**
>
> *— Sector: Non-profit; Tenure: 7+ years*

Technology Overview

Quantum computing is an emerging field that leverages quantum mechanical phenomena, resulting in exponentially faster computation of various computing tasks. Quantum computers operate using quantum bits, or qubits, which (unlike the bits used today in classical computing) can store not just a zero or a one, but also a weighted combination of zero and one. This quality of superposition is critical for unlocking the capability for quantum computers to store — and process in parallel — exponentially more information than classical computers. Other key qualities of quantum mechanical phenomena that similarly unlock these capabilities include entanglement (information about one qubit allows for immediate determination of information about other qubits) and interference (superposition results in probabilistic waves of information, where these probabilities can amplify or cancel each other out in pursuit of a computational outcome or measurement[282]). Quantum computing requires the use of specialized quantum hardware (much of which is currently experimental).

There are still concrete technical barriers and challenges to widespread use of quantum computing, chiefly its current unreliability in computing accurate results (also known as decoherence). Recent advances in logical qubits present promise for addressing that unreliability.[283,284] Today, some commercial applications of quantum computing already exist (e.g., its use in protein design[285,286] and for securing data).[287]

Current Opportunities & Impacts

Quantum computing remains largely in a theoretical state for application. Consequently, it has not yet been reported as playing a role either in the tools used to prevent and defend against technology-facilitated child sexual exploitation nor as a means for committing abuse. In the ecosystem survey, respondents with 3+ years of experience in the law enforcement sector* indicated that they were not fielding CSAM cases involving quantum computing, with all such respondents indicating that 0% of their CSAM cases involved this technology.

Future Opportunities & Impacts

While researchers and participants could not identify any current publicly reported cases involving quantum computing (at the time of publication), current theoretical work points to potential future risks and opportunities. Given its primary role (in theoretical work) as an accelerant of technologies, as opposed to a stand-alone service, it will to be helpful to consider the impact of quantum computing through that lens.

> **"QC can be used for a multitude of purposes for improving online child safety, whether thats identifying harmful content, or improving content moderation, and therefore makes these areas of work more efficient."**
>
> *— Sector: Other; Tenure: 7+ years*

Safety by Design

Resources advising on the use of quantum computing with child sexual abuse and exploitation harms in mind are scarce. Where that discussion is occurring, particular concerns around the use of quantum computing cyberattacks to break encryption algorithms have been noted.[288] Recently, the National Institute of Standards and Technology released a set of encryption tools designed to withstand these types of attacks.[289]

There appear to be two possible explanations for the lack of resources in this area. The first is that quantum computing is orthogonal to a broad range of computing tasks. In other words: quantum computing is, at its core, an accelerator, and so interventions to address its impact on technology-facilitated child sexual exploitation are better suited to sit within the specific task or computing workflow that is accelerated by the application of quantum computing. The second is that this is truly still an emerging technology, and ideating on impactful interventions for something that is still yet emerging can prove to be challenging.

> **"Quantum Computing could enable more rapid processing, uploading, etc., making it harder to prevent child safety risks before they occur."**
>
> *— Sector: Other; Tenure: 5-7 years*

*The base size of this group was small (n<25), and therefore these findings should be considered directional.

Discussion & Looking Ahead

Technology is evolving at a rapid pace. This yields vital tools for society that are at times even lifesaving. Invention has long required a careful examination of both the intended and unintended outcomes of what we create, and it has required pursuing necessary mitigations to avoid accidental harm stemming from something designed to benefit. Mitigations stem both from features embedded in the design as well as from the policies to oversee use of the tools and user instructions for safe use of the tool — balancing the likelihood of a negative impact by reducing capacity for harm and increasing awareness among users.

However, the role of user awareness has far greater limitations when examining the risk of unintended outcomes stemming from the intersection of technology and child safety, for two core reasons.

First, when the user is a minor, the ability to both educate on the nature and strategies to mitigate for unintended consequences is more limited than with adults, and at times purely beyond reasonable expectations. Further, for users in a developmental period where they are wired to take more risks and less likely to follow the rules and guardrails laid out for them by adults, the efficacy of such warnings is reduced.

Second, when the user is an adult, the outcomes we seek to prevent may in fact be the desired outcomes of someone intent on exploiting or abusing a minor. The potential to subvert product protections, violate terms of service, and manipulate other users is high.

For these reasons, the responsibility to balance user safety and innovation must include extra care in scrutinizing mitigation tactics for their likely efficacy.

This report outlines several areas of emerging or evolving technical innovation where work is needed to understand and address the risk of misuse for the purpose of child sexual abuse and exploitation. It does not focus on one singular technology over all others. Offenders have a long history of working across multiple surfaces to groom victims, network among offenders, and traffic in child sexual abuse material. **As a result, to focus on one at the exclusion of others will only further the game of "whack-a-mole," which we have played for the last several decades.**

This report also does not suggest that the solution to this problem is the elimination of these technologies nor the isolation of minors from the internet. Both digital innovation and connection are vital parts of society at this point; **attempts to deny young people access to the internet will restrict the real benefits and opportunities they offer to children and fail to support their successful onboarding to digital environments they will certainly utilize as they mature.**

Finally, the report not only lays out the current and future areas of potential abuse, but also highlights the importance of some of these technologies in delivering safer futures for young people. **The sheer volume of interactions online each day requires partnership between humans and technology. Neither can deliver without the other.** The efficacy of this partnership will require specialized subject matter and technical expertise paired with the will of the public and market to confront these risks.

Key areas of focus as we build to ensure that technology is designed and deployed with children in mind include:

Deepening and expanding the research base

Additional research to study the nature and impact of certain technologies on technology-facilitated child sexual exploitation is needed. Existing research is limited on many of these topics. As a result, we are often using blunt strategies targeting access and innovation over targeted interventions built upon strong data backing. Particular areas of focus should consider:

- Pursuing platform-agnostic initiatives to examine individual feature roles in user risk and safety.
- Investing in research among diverse populations geographically, culturally, economically, and developmentally to understand the impacts (both positive and negative) of technologies for different communities.
- Applying to child safety what has been learned in promising academic work on other areas of digital risk.

Multidisciplinary collaboration

Collaboration across disciplines is necessary to incorporate the knowledge of those on the front lines (investigators, survivor organizations, and nonprofits, to name a few) into the life cycle of technology development, deployment, and maintenance to ensure that builders have a strong understanding of how abuse of their technologies may occur. In addition, we should seek opportunities to learn from and collaborate with expertise not specific to technology-facilitated child sexual exploitation — for example, technologists, mis- and disinformation experts, and child learning and development — to apply core concepts from these fields to the more specialized field of technology-facilitated child sexual exploitation.

Safety by design frameworks

Too often, the question "How does this technology negatively impact children?" is asked after harm has occurred. Incorporation of dedicated safety by design workstreams during early exploration and design phases grounds child safety as foundational to the value and success of new technologies. For many of the technologies explored in this report, we are faced with considering how to contain these harms, rather than how to prevent the tools we build from being readily weaponized to cause harm in the first place. Changing this will require investment in safety by design efforts that are technologically sound, research based and multidisciplinary throughout the entire life cycle of technology.

Regular horizon scanning

High-level horizon scanning, focused on a rapid assessment of current and emerging technologies that are creating (or poised to create) disruption in the existing landscape of technology-facilitated child sexual exploitation, should be conducted on a regular cadence. This work, while challenging to prioritize, is critical to reduce risk to children in new technologies and nimbly respond to unintended impacts of technology.

Raising public awareness

Continuing to increase public awareness about the scale and nature of technology-facilitated child sexual exploitation is vital to ensuring we have not only the technical capabilities to combat this issue, but the will to prioritize and deploy them. Further, understanding the unique and critical expertise and roles of the many contributors across the ecosystem will lead to more holistic and effective strategies designed with children at the heart.

References

1. UNICEF. (2020). *Children and young people: Internet access at home during COVID-19*. https://data.unicef.org/resources/children-and-young-people-internet-access-at-home-during-covid19

2. UNICEF. (2023). *How many children under 18 are in the world?* https://data.unicef.org/how-many/how-many-children-under-18-are-in-the-world

3. eSafety Commissioner. (n.d.). *Safety by design.* https://www.esafety.gov.au/industry/safety-by-design

4. Thorn. (2024). *Generative AI principles*. https://www.thorn.org/blog/generative-ai-principles

5. Statista. (n.d.). *Artificial intelligence—Machine learning worldwide*. https://www.statista.com/outlook/tmo/artificial-intelligence/machine-learning/worldwide

6. Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156–180. https://doi.org/10.1109/RBME.2020.3013489

7. Ngai, E., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, 145, 35–48. https://doi.org/10.1016/j.jbusres.2022.02.049

8. Pattnaik, D., Ray, S., & Raman, R. (2024). Applications of artificial intelligence and machine learning in the financial services industry: A bibliometric review. *Heliyon,* 10(1), e23492. https://www.cell.com/heliyon/fulltext/S2405-8440(23)10700-6

9. Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems*, 74, 12–32. https://doi.org/10.1016/j.dss.2015.03.008

10. High-Level Committee on Programmes. (2022). *Principles for the ethical use of artificial intelligence in the United Nations System*. Inter-Agency Working Group on Artificial Intelligence, United Nations Systems. https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf

11. Future of Life Institute. (2024). *The EU Artificial Intelligence Art: Up-to-date developments and analyses of the EU AI Act*. https://artificialintelligenceact.eu

12. Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

13. Apodaca, T., & Uzcátegui-Liggett, N. (2024, March 1). How automated content moderation works (even when it doesn't). *The Markup*. https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work

14. Reyes, M. (2023, December 4). This is how artificial intelligence helps detect child sexual abuse material online. *Medium.* https://medium.com/@martareyessuarez25/this-is-how-artificial-intelligence-helps-detect-child-sexual-abuse-material-online-57aeed4d5463

15. Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

16. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement. Australian Government.* https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

17. Thiel, D., DiResta, R., and Stamos, A. (2023). Cross-platform dynamics of self-generated CSAM. *Stanford Digital Repository.* https://purl.stanford.edu/jd797tp7663

18. Horowitz, J., & Blunt, K. (2023, June 7). Instagram connects vast pedophile network. *The Wall Street Journal*. https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189

19. Dangerous by Design. (2021, December 8). "Thinstagram": Instagram's algorithm fuels eating disorder epidemic. *Tech Transparency Project.* https://www.techtransparencyproject.org/articles/thinstagram-instagrams-algorithm-fuels-eating-disorder-epidemic

20. Fathom. (2024, September). *AI at the crossroads: Public sentiment and policy solutions.* https://fathom.org/pdf/Report-AI-at-a-Crossroads.pdf

21. Government of the United Kingdom. (2023, November). *The Bletchley declaration by countries attending the AI Safety Summit, 1-2 November 2023*. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

22. Ryan-Mosley, T. (2023, May 15). Catching bad content in the age of AI. *MIT Technology Review*. https://www.technologyreview.com/2023/05/15/1073019/catching-bad-content-in-the-age-of-ai

23. European Commission. (2024). *Out-of-court dispute settlement bodies under the Digital Services Act (DSA)*. https://digital-strategy.ec.europa.eu/en/policies/dsa-out-court-dispute-settlement

24. Field, H., & Vanian, J. (2023, May 26). T*ech layoffs ravage the teams that fight online misinformation and hate speech*. CNBC. https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html

25. Basu, S. (2024, October 14). ByteDance lays off hundreds as TikTok shifts toward AI content moderation. *Readwrite.* https://readwrite.com/bytedance-lays-off-hundreds-tiktok-ai-content-moderation

26. Hanaoka, K., Ngen, M., Yang, J., Quinn, G., Hom, A., & Grother, P. (2024). *Face analysis technology evaluation: Age estimation and verification.* NIST Interagency Report 8525. National Institute of Standards and Technology, U.S. Department of Commerce. https://pages.nist.gov/frvt/reports/aev/fate_aev_report.pdf

27. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement. Australian Government*. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

28. Stray et al. (2024). *Building human values into recommender systems: An interdisciplinary synthesis.* https://arxiv.org/pdf/2207.10192

29. Ofcom. (2024). *Protecting children from harms online.* Consultation. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol5-what-should-services-do-to-mitigate-risks.pdf

30. Stray, J., Vendrov, I., Nixon, J., Adler, S.. & Hadfield-Menell, D. (2021). *What are you optimizing for?: Aligning recommender systems with human values.* https://arxiv.org/pdf/2107.10939

31. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement.* Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

32. Stray et al. (2024). *Building human values into recommender systems: An interdisciplinary synthesis*. https://arxiv.org/pdf/2207.10192

33. Stray, J., Vendrov, I., Nixon, J., Adler, S.. & Hadfield-Menell, D. (2021). *What are you optimizing for?: Aligning recommender systems with human values.* https://arxiv.org/pdf/2107.10939

34. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement.* Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

35. Meßmer, A., & Degeling, M. (2023). *Auditing recommender systems—Putting the DSA into practice with a risk-scenario-based approach*. Interface. https://arxiv.org/pdf/2302.04556

36. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement.* Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

37. Stray et al. (2024). *Building human values into recommender systems: An interdisciplinary synthesis.* https://arxiv.org/pdf/2207.10192

38. Thiel, D., DiResta, R., and Stamos, A. (2023). Cross-platform dynamics of self-generated CSAM. *Stanford Digital Repository.* https://purl.stanford.edu/jd797tp7663

39. Stray, J., Vendrov, I., Nixon, J., Adler, S.. & Hadfield-Menell, D. (2021). *What are you optimizing for?: Aligning recommender systems with human values.* https://arxiv.org/pdf/2107.10939

40. O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. *Proceedings of the 2005 International Conference on Intelligent User Interfaces.* https://www.researchgate.net/publication/221608315_Trust_in_recommender_systems

41. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement.* Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

42. Ofcom. (2024). *Protecting children from harms online.* Consultation. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol5-what-should-services-do-to-mitigate-risks.pdf

43. eSafety Commissioner. (2023). *Recommender systems and algorithms: Position statement.* Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms/full-position-statement

44. Ofcom. (2024). *Protecting children from harms online.* Consultation. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol5-what-should-services-do-to-mitigate-risks.pdf

45. Thiel, D., DiResta, R., and Stamos, A. (2023). Cross-platform dynamics of self-generated CSAM. *Stanford Digital Repository.* https://purl.stanford.edu/jd797tp7663

46. Digital Services Act, articles 26–28, 38. https://www.eu-digital-services-act.com/Digital_Services_Act_Articles.html

47. European Union Artificial Intelligence Act, article 5 (1a,b). https://artificialintelligenceact.eu/article/5

48. Gozalo-Brizuela, R., & Garrido-Merchàn, E. (2023). *A survey of generative AI applications.* https://arxiv.org/pdf/2306.02781

49. Statista. (n.d.). *Generative AI—Worldwide*. https://www.statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide

50. United Nations. (2023, July 18). *International community must urgently confront new reality of generative, artificial intelligence, speakers stress as Security Council debates risks, rewards.* https://press.un.org/en/2023/sc15359.doc.htm

51. Ferri, G., & Gloerich, I. (2023). Risk and harm: Unpacking ideologies in the AI discourse. *Proceedings of the 5th International Conference on Conversational User Interfaces*, article 28. https://dl.acm.org/doi/abs/10.1145/3571884.3603751

52. Ravaglia, R. (2023, August 28). Generative AI makes Grammarly an essential student learning tool. *Forbes.* https://www.forbes.com/sites/rayravaglia/2023/08/28/generative-ai-makes-grammarly-an-essential-student-learning-tool

53. Mathews, A. (2024). *The future of music is in AI: Thanks to Spotify.* AIM Research. https://aimresearch.co/market-industry/the-future-of-music-is-in-ai-thanks-to-spotify

54. GE Appliances. (2023, August 29). *GE Appliances helps consumers create personalized recipes from the food in their kitchen with Google Cloud's generative AI.* https://pressroom.geappliances.com/news/ge-appliances-helps-consumers-create-personalized-recipes-from-the-food-in-their-kitchen-with-google-clouds-generative-ai

55. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

56. Thorn. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

57. Internet Watch Foundation. (2024, July). *What has changed in the AI CSAM landscape?* AI SCAM report update. https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf

58. Saliba, E. (2023, July 28). *Sharing photos of your kids? Maybe not after you watch this deepfake ad.* ABC News. https://abcnews.go.com/GMA/Family/sharing-photos-kids-after-watch-deepfake-ad/story?id=101730561

59. Brewster, T. (2024, August 30). Pedophile filmed kids at Disney World to make AI child abuse images cops say. Forbes. https://www.forbes.com/sites/thomasbrewster/2024/08/30/pedophile-filmed-kids-at-disney-world-to-make-ai-child-abuse-images-cops-say

60. Aronashvili, R. (2024, February 20). The evolution of sextortion attacks: How generative AI is taking a front seat. *Forbes.* https://www.forbes.com/councils/forbestechcouncil/2024/02/20/the-evolution-of-sextortion-attacks-how-generative-ai-is-taking-a-front-seat

61. Thorn and National Center for Missing and Exploited Children (NCMEC). (2024). *Trends in financial sextortion: An investigation of sextortion reports in NCMEC CyberTipline data*. https://www.thorn.org/research/library/financial-sextortion

62. Jargon, J. (2023, November 2). Fake nudes of real students cause an uproar at a New Jersey high school. *The Wall Street Journal.* https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb?st=4qmspwzwhxlcdsr&reflink=desktopwebshare_permalink&mod=article_inline

63. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

64. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B. B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

65. Thiel, D. (2023). Identifying and eliminating CSAM in generative ML training data and models. *Stanford Digital Repository*. https://doi.org/10.25740/kh752sm9123

66. Steinberg, S. (2019). Changing faces: Morphed child pornographhy images and the First Amendment. *Emory Law Journal, 68* (5). https://scholarlycommons.law.emory.edu/cgi/viewcontent.cgi?article=1025&context=elj

67. Lanning, K. (2010). *Child molesters: A behavioral analysis, 5th ed.* National Center for Mission & Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/ publications/nc70.pdf

68. Thorn. (2024). *Youth perspectives on online safety, 2023.* https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf

69. Singer, N. (2024, April 8). Teen girls confront an epidemic of deepfake nudes in schools. *The New York Times.* https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html

70. Guy, J. (2023, September 20). *Outcry in Spain as artificial intelligence used to create fake naked images of underage girls.* CNN. https://www.cnn.com/2023/09/20/europe/spain-deepfake-images-investigation-scli-intl/index.html

71. Healey, J. (2024, February 26). Beverly Hills middle school rocked by AI-generated nude images of students. *Los Angeles Times.* https://www.latimes.com/california/story/2024-02-26/beverly-hills-middle-school-is-the-latest-to-be-rocked-by-deepfake-scandal

72. Haskins, C. (2024, March 8). Florida middle schoolers arrested for allegedly creating depfake nudes of classmates. *Wired*. https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates

73. Jargon, J. (2023). Fake nudes of real students cause an uproar at a New Jersey high school. *The Wall Street Journal.* https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb?st=4qmspwzwhxlcdsr&reflink=desktopwebshare_permalink&mod=article_inline

74. UN Committee on the Rights of the Child. (2021). General comment No. 25 on children's rights in relation to the digital environment (CRC/C/GC/25). https://documents.un.org/doc/undoc/gen/g21/053/43/pdf/g2105343.pdf

75. Development Services Group. (2017). *Diversion from formal juvenile court processing: Literature review.* Office of Juvenile Justice and Prevention, U.S. Department of justice. https://ojjdp.ojp.gov/model-programs-guide/literature-reviews/diversion_from_formal_juvenile_court_processing.pdf

76. Grossman, S., Pfefferkorn, R., Thiel, D., Shah, S., DiResta, R., Perrino, J., Cryst, E., Stamos, A., & Hancock, J. (2024). The strengths and weaknesses of the online child safety ecosystem. *Stanford Digital Repository.* https://purl.stanford.edu/pr592kc5483. https://doi.org/10.25740/pr592kc5483

77. Popken, A. (2024, June 7). We're unprepared for the threat GenAI on Instagram, Facebook, and Whatsapp poses to kids. *Fast Company.* https://www.fastcompany.com/91136311/were-unprepared-for-the-threat-genai-on-instagram-facebook-and-whatsapp-poses-to-kids

78. Associated Press. (2017, March 27). *"Swirl face" sexual offender in Vancouver after release.* BBC News. https://www.bbc.com/news/world-us-canada-39411025

79. Lucy Faithfull Foundation. (2024, September 12). *Introducing our chatbot—A pioneering tool in the fight against online child sexual abuse*. https://www.lucyfaithfull.org.uk/featured-news/Introducing_our_chatbot_a_pioneering_tool.htm

80. Smith, S. (2024, September 12). *Chatbots and warning messages: Innovations in the fight against online child sexual abuse.* Lucy Faithfull Foundation. https://www.lucyfaithfull.org.uk/files/Faithfull_Paper_Chatbots_12_Sept_2024.pdf

81. IBM. (n.d.). *What are AI hallucinations?*. https://www.ibm.com/topics/ai-hallucinations

82. Xiang, C. (2023, March 30). "He would still be here": Man dies by suicide after talking with AI chatbot, widow says. *Vice.* https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

83. Goudsward, A. (2024, October 19). *US prosecutors see rising threat of AI-generated child sex abuse imagery.* Reuters. https://www.reuters.com/technology/us-prosecutors-see-rising-threat-ai-generated-child-sex-abuse-imagery-2024-10-17

84. Ng, F. (May 24, 2023). *AI depfakes are getting better at spoofing KYC verification—Binance exec.* Cointelegraph. https://cointelegraph.com/news/binance-rise-in-deepfake-customer-checks-verification

85. Notopolis, K. (2024, October). Harvard students use Meta Ray Ban to demo facial recognition. *Business Insider.* https://www.businessinsider.com/meta-ray-ban-glasses-facial-recognition-demo-students-2024-10

86. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse*. Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

87. National Institute of Standards and Technology. *Reducing risks posed by synthetic content.* NIST AI 100-4 Draft for public comment. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

88. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes*. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

89. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

90. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

91. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

92. UNICEF. (2021, November). *Policy guidance on AI for children, version 2.* https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

93. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse*. Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

94. National Institute of Standards and Technology. *Reducing risks posed by synthetic content.* NIST AI 100-4 Draft for public comment. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

95. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

96. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

97. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

98. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

99. National Institute of Standards and Technology. *Reducing risks posed by synthetic content.* NIST AI 100-4 Draft for public comment. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

100. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

101. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

102. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

103. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

104. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

105. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

106. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

107. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

108. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

109. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

110. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

111. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

112. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

113. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

114. UNICEF. (2021, November). *Policy guidance on AI for children, version 2.* https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

115. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

116. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

117. *EU Artificial Intelligence Act, Recital 133.* Future of Life Institute. https://artificialintelligenceact.eu/recital/133

118. *EU Artificial Intelligence Act, Recital 134.* Future of Life Institute. https://artificialintelligenceact.eu/recital/134

119. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.* Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

120. National Institute of Standards and Technology. *Reducing risks posed by synthetic content.* NIST AI 100-4 Draft for public comment. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

121. National Institute of Standards and Technology. (2024, July). *Artificial intelligence risk management framework: Generative artificial intelligence profile.* NIST AI 600-1. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

122. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

123. Srikumar, M., Chmielinski, K., Chang, J. (2024, July). *Risk mitigation strategies for the Open Foundation Model value chain.* Partnership on AI. https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf

124. Thiel, D., Stroebel, M., & Portnoff, R. (2023). *Generative ML and CSAM: Implications and mitigations.* Thorn and Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

125. Portnoff, R., Simpson, M., Wang, R., O'Gorman, T., Polgar, D., Tweed, R., Cummings, R., Hallinan, P., Myers, M., Schaffer, M., Herrera, S., Issembert, B.B., Brooks, B., & Lalani, F. (2024). *Safety by design for generative AI: Preventing child sexual abuse.*

126. Internet Watch Foundation. (2023, October). *How AI is being abused to create child sexual abuse imagery.* https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf

127. National Institute of Standards and Technology. *Reducing risks posed by synthetic content.* NIST AI 100-4 Draft for public comment. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

128. Ofcom (2024, July 23). *Deepfake defences: Mitigating the harms of deceptive deepfakes.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

129. eSafety Commissioner. (2023). *Generative AI—Tech trends position statement.* https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

130. UNICEF. (2021, November). *Policy guidance on AI for children, version 2.* https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

131. Apple. (2024, September 16). *iCloud data security overview.* https://support.apple.com/en-us/102651

132. 1Password Support. (2022, October 5). *Authentication and encryption in the 1Password security model.* https://support.1password.com/authentication-encryption

133. Stripe Support. (n.d.). *Stripe terminal encryption: E2EE vs. P2PE.* https://support.stripe.com/questions/stripe-terminal-encryption-e2ee-vs-p2pe

134. National Center for Missing & Exploited Children. (2023). *CyberTipline report.* https://www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf

135. National Center for Missing & Exploited Children. (2023). *CyberTipline report.* https://www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf

136. Thorn. (2024, July 19). *Introducing Safer Predict: Using the power of AI to detect child sexual abuse and exploitation online.* https://www.thorn.org/blog/introducing-safer-predict-using-the-power-of-ai-to-detect-child-sexual-abuse-and-exploitation-online

137. Google. (n.d.). *Tools for partners*. https://protectingchildren.google/tools-for-partners

138. Business for Social Responsibility. (n.d.). *Meta's expansion of End-to-End Encryption: Human Rights Impact Assessment*. https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf

139. Meta. (2022, April). *Meta response: End-to-end encryption human rights impact assessment*. https://about.fb.com/wp-content/uploads/2022/04/E2EE-HRIA-Meta-Response.pdf

140. BSR. (n.d.). *Meta's expansion of End-to-End Encryption: Human Rights Impact Assessment.* https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf

141. Winters, G., Colombino, N., Schaaf, S., Laake, A., Jeglic, E., & Calkins, C. (2020). Why do child sexual abuse victims not tell anyone about their abuse? An exploration of factors that prevent and promote disclosure. *Behavioral Sciences & the Law, 38*(6). https://doi.org/10.1002/bsl.2492

142. Lahtinen, H., Laitila, A., Korkman, J., & Ellonen, N. (2018). Children's disclosures of sexual abuse in a population-based sample. *Child Abuse & Neglect, 76*, 84–94. https://doi.org/10.1016/j.chiabu.2017.10.011

143. Thorn. (2024, August). *Youth perspectives on online safety, 2023.* https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf

144. Thorn. (2021, May). *Responding to online threats: Minors' perspectives on disclosing, reporting, and blocking.* https://info.thorn.org/hubfs/Research/Responding%20to%20Online%20Threats_2021-Full-Report.pdf

145. Hatmaker, T. (2021, November 30). *Twitch's newest moderation tool spots accounts trying to get around channel bans.* TechCrunch. https://techcrunch.com/2021/11/30/twitch-flag-banned-users-from-channelc

146. Thorn. (2021, May). *Responding to online threats: Minors' perspectives on disclosing, reporting, and blocking.* https://info.thorn.org/hubfs/Research/Responding%20to%20Online%20Threats_2021-Full-Report.pdf

147. Darkness to Light. (2020, January 29). *Grooming and Red Flag Behaviors.* https://www.d2l.org/child-grooming-signs-behavior-awareness/

148. ECPAT International. (2024, May 15). *ECPAT'S SUBMISSION Call for input: Existing and Emerging Sexually Exploitative Practices against Children in the Digital Environment.* https://www.ohchr.org/sites/default/files/documents/issues/children/sr/cfis/existing-emerging/subm-existing-emerging-sexually-cso-ecpat.pdf

149. Thorn. (2022, April). *Online grooming: examining risky encounters amid everyday digital socialization.* https://info.thorn.org/hubfs/Research/2022_Online_Grooming_Report.pdf

150. Thorn. (2024, June). *Trends in financial sextortion.* https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf

151. Crisan, L. (2023, December 6). *Launching default end-to-end encryption on Messenger.* Meta. https://about.fb.com/news/2023/12/default-end-to-end-encryption-on-messenger

152. National Center for Missing & Exploited Children. (2023, December 7). *A devastating blow to child protection: Meta expands encryption.* https://www.missingkids.org/blog/2023/devastating-blow-child-protection-meta-expands-encryption

153. National Center for Missing & Exploited Children. (2023). *2023 CyberTipline reports by electronic service providers (ESP).* https://www.missingkids.org/content/dam/missingkids/pdfs/2023-reports-by-esp.pdf

154. Abelson H., Anderson, R., Bellovin, S., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P., Rivest, R., Schiller, J., Schneier, B., Teague, V., & Troncoso, C. (2021, October 15). *Bugs in our pockets: The risks of client-side scanning.* https://arxiv.org/pdf/2110.07450

155. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms.* https://arxiv.org/pdf/2207.09506

156. eSafety Commissioner. (2023). *Updated position statement—End-to-end encryption.* Australian Governemnet. https://www.esafety.gov.au/sites/default/files/2023-10/End-to-end-encryption-position-statement-oct2023.pdf

157. Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021, August). *Outside looking in: Approaches to content moderation in end-to-end encrypted systems*. https://arxiv.org/pdf/2202.04617

158. Pfefferkorn, R. (2022). Content-oblivious trust and safety techniques: Results from a survey of online service providers. *Journal of Online Trust and Safety*. https://doi.org/10.54501/jots.v1i2.14

159. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms*. https://arxiv.org/pdf/2207.09506

160. Scheffler, S., & Mayer, J. (2023). *SoK: Content moderation for end-to-end encryption.* https://arxiv.org/pdf/2303.03979

161. eSafety Commissioner. (2023). *Updated position statement—End-to-end encryption.* Australian Governemnet. https://www.esafety.gov.au/sites/default/files/2023-10/End-to-end-encryption-position-statement-oct2023.pdf

162. Scheffler, S., & Mayer, J. (2023). *SoK: Content moderation for end-to-end encryption.* https://arxiv.org/pdf/2303.03979

163. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms.* https://arxiv.org/pdf/2207.09506

164. eSafety Commissioner. (2023). *Updated position statement—End-to-end encryption.* Australian Governemnet. https://www.esafety.gov.au/sites/default/files/2023-10/End-to-end-encryption-position-statement-oct2023.pdf

165. Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021, August). *Outside looking in: Approaches to content moderation in end-to-end encrypted systems*. https://arxiv.org/pdf/2202.04617

166. Tyagi, N., Grubbs, P., Len, J., Miers, I., & Ristenpart, T. (2019). Asymmetric message franking: Content moderation for metadata–private end-to-end encryption. Conference paper in *Advances in Cryptology—CRYPTO 2019, 11694*, 222–250. https://doi.org/10.1007/978-3-030-26954-8_8

167. Scheffler, S., & Mayer, J. (2023). *SoK: Content moderation for end-to-end encryption*. https://arxiv.org/pdf/2303.03979

168. Scheffler, S., Kulshrestha, A., & Mayer, J. (2023). *Public verification for private hash matching.* International Association for Cryptologic Research. https://eprint.iacr.org/2023/029.pdf

169. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms*. https://arxiv.org/pdf/2207.09506

170. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms.* https://arxiv.org/pdf/2207.09506

171. Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021, August). *Outside looking in: Approaches to content moderation in end-to-end encrypted systems.* https://arxiv.org/pdf/2202.04617

172. Pfefferkorn, R. (2022). Content-oblivious trust and safety techniques: Results from a survey of online service providers. *Journal of Online Trust and Safety.* https://doi.org/10.54501/jots.v1i2.14

173. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms.* https://arxiv.org/pdf/2207.09506

174. Kulshrestha, A., & Mayer, J. (2021). Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation. *Proceedings of the 30th USENIX Security Symposium,* 893–910. https://www.usenix.org/system/files/sec21-kulshrestha.pdf

175. Acar, A., Aksu, H., & Uluagac, A. S. (2017). *A survey on homomorphic encryption schemes: Tehory and implementation.* https://arxiv.org/pdf/1704.03578

176. Levy, I., & Robinson, C. (2022, July 21). *Thoughts on child safety on commodity platforms.* https://arxiv.org/pdf/2207.09506

177. Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021, August). *Outside looking in: Approaches to content moderation in end-to-end encrypted systems.* https://arxiv.org/pdf/2202.04617

178. Abelson H., Anderson, R., Bellovin, S., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P., Rivest, R., Schiller, J., Schneier, B., Teague, V., & Troncoso, C. (2021, October 15). *Bugs in our pockets: The risks of client-side scanning*. https://arxiv.org/pdf/2110.07450

179. PokemonGo. (n.d.). https://pokemongolive.com; Meta. (n.d.). *Get started with Meta Quest 2.* https://www.meta.com/quest/products/quest-2

180. Virtual Medicine. (n.d.). *The most advanced VR anatomy platform.* https://www.medicinevirtual.com; Google. (n.d.). *Google Lens.* https://lens.google; IKEA. (2017, September 12)

181. *IKEA Place app launched to help people virtually place furniture at home.* https://www.ikea.com/global/en/newsroom/innovation/ikea-launches-ikea-place-a-new-app-that-allows-people-to-virtually-place-furniture-in-their-home-170912

182. Ford. (2017, September 21). *Make way for holograms: New mixed reality technology meets car design as Ford tests Microsoft HoloLens globally.* https://media.ford.com/content/fordmedia/fna/us/en/news/2017/09/21/ford-tests-microsoft-hololens-globally.html

183. Chandukala, S., Reddy, S., & Tan, Y. (2022, March 29). How augmented reality can—and can't—help your brand. *Harvard Business Review.* https://hbr.org/2022/03/how-augmented-reality-can-and-cant-help-your-brand

184. Marr, B. (2021, December 10). 10 best examples of VR and AR in education. *Forbes.* https://www.forbes.com/sites/bernardmarr/2021/07/23/10-best-examples-of-vr-and-ar-in-education

185. Cantor, C. (2022, July 27). *Virtual reality helps teens and young adults with social anxiety.* Columbia University Department of Psychiatry. https://www.columbiapsychiatry.org/news/virtual-reality-can-help-teens-and-young-adults-social-anxiety

186. Spencer, T. (2024, February 19). *"Soaring" over hills or "playing" with puppies, study finds seniors enjoy virtual reality.* Associated Press. https://apnews.com/article/virtual-reality-seniors-florida-dementia-c2989fcfb5dca58639fbc0f8060d1eff

187. https://www.businesswire.com/news/home/20210822005004/en/Juniper-Research-Global-Revenue-from-Adult-Virtual-Reality-Content-to-Reach-19-Billion-by-2026-as-Subscription-Models-Dominate

188. India Today Tech. (2022, June 22). VR porn search sees a spike by 115 per cent in five months. *India Today.* https://www.indiatoday.in/technology/news/story/vr-porn-sees-a-spike-in-search-by-115-per-cent-in-five-months-1965407-2022-06-22

189. Vallance, C. (2024, January 2). *Police investigate virtual sex assault on girl's avatar.* BBC. https://www.bbc.com/news/technology-67865327

190. Crawford, A. (2023, February 22). *Child abuse material found on VR headsets, police data shows.* BBC. https://www.bbc.com/news/uk-64734308

191. Allen, C. (n.d.). *Child safeguarding and immersive technologies: An outline of the risks.* National Society for the Prevention of Cruelty to Children. https://learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf

192. Hart, K. (2024, August 1). Man charged after child says she was sexually abused while playing VR. *East Idaho News.* https://www.eastidahonews.com/2024/08/man-charged-after-child-says-she-was-sexually-abused

193. Roth, E. (2024, July 10). Meta will soon let young kids chat in VR—But only with their parents' approval. *The Verge.* https://www.theverge.com/2024/7/10/24195692/meta-quest-approved-contacts-parental-controls

194. Crawford, A., & Smith, T. (2022, February 23). *Metaverse app allows kids into virtual strip clubs.* BBC. https://www.bbc.com/news/technology-60415317

195. Bark. (2024, September 12). *How to set up Meta Quest parental controls*. Tech guides. https://www.bark.us/tech-guide/gaming-oculus/?srsltid=AfmBOoofK-LuHqVSNs7rLMiRCnTH1rQuggdeWewG-tWzOhEV0jFImPIA

196. Allen, C. (n.d.). *Child safeguarding and immersive technologies: An outline of the risks.* National Society for the Prevention of Cruelty to Children. https://learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf

197. Crawford, A., & Smith, T. (2022, February 23). *Metaverse app allows kids into virtual strip clubs.* BBC. https://www.bbc.com/news/technology-60415317

198. Cross, R. (2023, December 6). *VR risks for kids and teens.* U.S. Public Interest Research Group Education Fund. https://pirg.org/edfund/resources/vr-risks-for-kids

199. Jiang, J., Kiene, C., Middler, S., Brubaker, J., & Fiesler, C. (2019). Moderation challenges in voice-based online communities on Discord. *Proceedings of the Association for Computing Machinery on Human-Computer Interaction, 3*, Article 55. https://doi.org/10.1145/3359157

200. Sabri, N., Teoh, A., Vaccaro, K., Chen, B., Dow, S., & ElSherief, M. (2023). Challenges of moderating social virtual reality. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 23–28, 2023, Hamburg, Germany.

201. Blackwell, L., Ellison, N., Elliott-Deflo, N., & Schwartz, R. (2019). Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the Association for Computing Machinery on Human-Computer Interaction, 3,* Article 100. https://doi.org/10.1145/3359202

202. Belamire, J. (2016, October 20). My first virtual reality groping. *Medium.* https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee

203. Sum of Us. (2022, May). *Metaverse: Another cesspool of toxic content.* https://www.eko.org/images/Metaverse_report_May_2022.pdf

204. Patel, N. (2021, December 21). Reality of fiction? *Medium*. https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0

205. Belamire, J. (2016, October 20). My first virtual reality groping. *Medium.* https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee

206. Blackwell, L., Ellison, N., Elliott-Deflo, N., & Schwartz, R. (2019). Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the Association for Computing Machinery on Human-Computer Interaction, 3*, Article 100. https://doi.org/10.1145/3359202

207. Tang-Lippert, K. (2024, January). A sexual assault in the metaverse has investigators questioning the future of virtual reality. *Business Insider.* https://www.businessinsider.com/police-investigate-digital-gang-rape-teen-vr-metaverse-horizon-worlds-2024-1

208. Camber, R. (2024, January 1). British police probe virtual rape in metaverse: Young girl's digital persona "is sexually attacked by gang of adult men in immersive video game"—sparking first investigation of its kind and questions about extent current laws apply in online world. *Daily Mail*. https://www.dailymail.co.uk/news/article-12917329/Police-launch-investigation-kind-virtual-rape-metaverse.html

209. Nix, N. (2024, February 6). Attacks in the metaverse are booming. Police are starting to pay attention. *Washington Post.* https://www.washingtonpost.com/technology/2024/02/04/metaverse-sexual-assault-prosecution

210. Sparrow, L., Antonello, M., Gibbs, M., & Arnold, M. (2020). From "silly" to "scumbag": Reddit discussion of a case of groping in a virtual reality game. *Proceedings of Digital Games and Research Association.* https://rest.neptune-prod.its.unimelb.edu.au/server/api/core/bitstreams/347bc7a3-1297-5078-a286-edd95e1d448e/content

211. Sparrow, L., Antonello, M., Gibbs, M., & Arnold, M. (2020). From "silly" to "scumbag": Reddit discussion of a case of groping in a virtual reality game. *Proceedings of Digital Games and Research Association.* https://rest.neptune-prod.its.unimelb.edu.au/server/api/core/bitstreams/347bc7a3-1297-5078-a286-edd95e1d448e/content

212. Allen, C. (n.d.). *Child safeguarding and immersive technologies: An outline of the risks.* National Society for the Prevention of Cruelty to Children. https://learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf

213. Allen, C. (n.d.). *Child safeguarding and immersive technologies: An outline of the risks.* National Society for the Prevention of Cruelty to Children. https://learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf

214. Thorbecke, C. (2023, March 3). *Meta is cutting prices for its VR headsets.* CNN. https://www.cnn.com/2023/03/03/tech/meta-vr-headsets-price-cut/index.html

215. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children*. https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

216. Zheng, Q., Xu, S., Wang, L., Tang, Y., Salvi, R., Freeman, G., & Huang, Y. (2023, April 16). Understanding safety risks and safety design in social VR environments. *Proceedings of the ACM on Human-Computer Interaction, 7,* 154. https://dl.acm.org/doi/abs/10.1145/3579630

217. Blackwell, L., Ellison, N., Elliott, N., & Schwartz, R. (2019). *Harassment in social virtual reality: Challenges for platform governance.* http://www.lindsayblackwell.net/wp-content/uploads/2019/09/Harassment-in-Social-Virtual-Reality-CSCW-2019.pdf

218. XR Safety Initiative. *The XRSI privacy framework*. https://xrsi.org/wp-content/uploads/2020/09/XRSI-Privacy-Framework-v1_002.pdf

219. Cortese, M., & Outlaw, J. (2021). *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report—Social and multi-user spaces in VR: Trolling, harassment, and online safety*. IEEE SA Industry Connections Report, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9650825

220. Responsible Metaverse Alliance, Wallace, C., Rosenberg, L., Pearlman, K., & Choudhary, B. (2023, May). *The metaverse and standards.* White paper. https://xrsi.org/wp-content/uploads/2023/07/646c3ffae5f92f9700dbe320_H2_3061_Metaverse_report.pdf

221. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children.* https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

222. Zheng, Q., Xu, S., Wang, L., Tang, Y., Salvi, R., Freeman, G., & Huang, Y. (2023, April 16). Understanding safety risks and safety design in social VR environments. *Proceedings of the ACM on Human-Computer Interaction, 7,* 154. https://dl.acm.org/doi/abs/10.1145/3579630

223. De Guzman, J., Thilakarathna, K., & Seneviratne, A. (2020). *Security and privacy approaches in mixed reality: A literature survey.* https://arxiv.org/pdf/1802.05797

224. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

225. Gray, J., Carter, M., & Egliston, B. (2024). Designing for safety, privacy and inclusivity in social VR. *Governing social virtual reality.* Springer Nature. https://www.springerprofessional.de/en/designing-for-safety-privacy-and-inclusivity-in-social-vr/27446820

226. Zytko, D., & Chan, J. (2023). The dating metaverse: Why we need to design for consent in social VR. *IEEE Transaction on Visualization and Computer Graphics.* https://dougzytko.com/research/Zytko_Chan_TVCG.pdf

227. Responsible Metaverse Alliance, Wallace, C., Rosenberg, L., Pearlman, K., & Choudhary, B. (2023, May). *The metaverse and standards.* White paper. https://xrsi.org/wp-content/uploads/2023/07/646c3ffae5f92f9700dbe320_H2_3061_Metaverse_report.pdf

228. Guzman, J., Thilakarathna, K., & Seneviratne, A. (2020). *Security and privacy approaches in mixed reality: A literature survey.* https://arxiv.org/pdf/1802.05797

229. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children.* https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

230. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

231. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

232. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children.* https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

233. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

234. Maloney, D., Freeman, G., & Robb, A. (2020). *A virtual space for all: Exploring children's experience in social virtual reality. CHI Play '20,* November 2–4, 2020, virtual, Canada. https://www.andrewrobb.io/publication/maloney-2020-virtual/maloney-2020-virtual.pdf

235. Responsible Metaverse Alliance, Wallace, C., Rosenberg, L., Pearlman, K., & Choudhary, B. (2023, May). *The metaverse and standards.* White paper. https://xrsi.org/wp-content/uploads/2023/07/646c3ffae5f92f9700dbe320_H2_3061_Metaverse_report.pdf

236. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children.* https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

237. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

238. Gray, J., Carter, M., & Egliston, B. (2024). *Designing for safety, privacy and inclusivity in social VR. Governing social virtual reality.* Springer Nature. https://www.springerprofessional.de/en/designing-for-safety-privacy-and-inclusivity-in-social-vr/27446820

239. Responsible Metaverse Alliance, Wallace, C., Rosenberg, L., Pearlman, K., & Choudhary, B. (2023, May). *The metaverse and standards.* White paper. https://xrsi.org/wp-content/uploads/2023/07/646c3ffae5f92f9700dbe320_H2_3061_Metaverse_report.pdf

240. Davidson, J., Farr, R., Bradbury, P., & Meggyesfalvi, B. (2024). *VIRRAC toolkit report: Virtual reality risks against children.* https://repository.uel.ac.uk/download/d7594f9a352fb69cf504f3d927b325a3a31587ccc04735b86dfd63b0b65d1d4a/2233994/VIRRAC%20Toolkit%20Report%202024%20%28Public%20Facing%20Summary%29.pdf

241. eSafety Commissioner (2024). *Immersive technologies—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2020-12/Immersive%20tech%20-%20Position%20statement.pdf

242. Cortese, M., & Outlaw, J. (2021). *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report—Social and multi-user spaces in VR: Trolling, harassment, and online safety.* IEEE SA Industry Connections Report, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9650825

243. Responsible Metaverse Alliance, Wallace, C., Rosenberg, L., Pearlman, K., & Choudhary, B. (2023, May). *The metaverse and standards.* White paper. https://xrsi.org/wp-content/uploads/2023/07/646c3ffae5f92f9700dbe320_H2_3061_Metaverse_report.pdf

244. Guzman, J., Thilakarathna, K., & Seneviratne, A. (2020). *Security and privacy approaches in mixed reality: A literature survey.* https://arxiv.org/pdf/1802.05797

245. Abhinaya, S. B., Sabir, A., & Das, A. (n.d.). *Enabling developers, protecting users: Investigating harassment and safety in VR.* https://www.usenix.org/system/files/sec24fall-prepub-329-sb.pdf

246. Allen, C. (n.d.). *Child safeguarding and immersive technologies: An outline of the risks.* National Society for the Prevention of Cruelty to Children. https://learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf

247. Mastodon. (n.d.). https://joinmastodon.org

248. IPFS. (n.d.). https://ipfs.tech

249. Bitcoin. (n.d.). https://bitcoin.org/en

250. IBM. (n.d.). *What is federated learning?* https://research.ibm.com/blog/what-is-federated-learning

251. https://www.britannica.com/technology/fediverse

252. Internet Watch Foundation. (2022, November 1). *Websites offering cryptocurrency payment for child sexual abuse images "doubling every year."* https://www.iwf.org.uk/news-media/news/websites-offering-cryptocurrency-payment-for-child-sexual-abuse-images-doubling-every-year

253. Greenberg, A. (2022, April 7). The crypto trap: Inside the Bitcoin bust that took down the web's biggest child abuse site. *Wired.* https://www.wired.com/story/tracers-in-the-dark-welcome-to-video-crypto-anonymity-myth

254. Chainalysis Team. (2024, January 11). *CSAM and cryptocurrency: On-chain analysis suggests CSAM vendors may benefit from privacy coins like Monero and other obfuscation measures.* https://www.chainalysis.com/blog/csam-cryptocurrency-monero-instant-exchangers-2024

255. Grant, J. (2021, July 30). Removing the risks from a decentralized internet. *The Strategist.* https://www.aspistrategist.org.au/removing-the-risks-from-a-decentralised-internet; Thiel, D., & DiResta, R. (2023, July 24). *Child safety on federated social media.* Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:vb515nd6874/20230724-fediverse-csam-report.pdf

256. Newton, C. (2019, December 16). The terror queue. *The Verge.* https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video

257. Pilling, D. (2023, May 17). "You can't unsee it": The content moderators taking on Facebook. *Financial Times*. https://www.ft.com/content/afeb56f2-9ba5-4103-890d-91291aea4caa

258. Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 17*(4), Article 8. https://doi.org/10.5817/CP2023-4-8

259. Concentrix. (2024, July 16). *Wellbeing programs: Protecting the mental health of content moderators*. https://www.concentrix.com/insights/blog/protecting-the-wellbeing-of-content-moderators

260. Thiel, D., & DiResta, R. (2023, July 24). *Child safety on federated social media.* Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:vb515nd6874/20230724-fediverse-csam-report.pdf

261. Thiel, D [@det]. (2023, July 23). Hits were primarily on a not-to-be-named Japanese instance, but a secondary test to see how far they propagated did show them . . . [Hachyderm post]. Mastodon. https://hachyderm.io/@det/110769472431035793

262. Goggin, B., Kolodny, L., CNBC, & Ingram, D. (2023, January 6). *On Musk's Twitter, users looking to sell and trade child sex abuse material are still easily found.* NBC News. https://www.nbcnews.com/tech/tech-news/musk-twitter-elon-child-abuse-material-rcna63621

263. Wang, F. (2024). Breaking the silence: Examining process of cyber sextortion and victims' coping strategies. *International Review of Victimology.* https://journals.sagepub.com/doi/pdf/10.1177/02697580241234331

264. Stamos, A., & Shah, S. (2023, October 17). *Common abuses on Mastodon: A primer.* Freeman Spogli Institute for International Studies. https://fsi.stanford.edu/news/common-abuses-mastodon-primer

265. Grant, J. (2023, October 16). Tech companies must do more to stem the tide of online child sexual abuse. *Tech Policy Press.* https://www.techpolicy.press/tech-companies-must-do-more-to-stem-the-tide-of-online-child-sexual-abuse/

266. Editorial Board (n.d.). The Durov case is not about free speech. *Financial Times.* https://www.ft.com/content/290a6308-58ab-45c8-b66f-d055373578c0

267. Thiel, D., & DiResta, R. *Child safety on federated social media.* Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:vb515nd6874/20230724-fediverse-csam-report.pdf

268. Hassan, A., Raman, A., Castro, I., Bin Zia, H., De Cristofaro, E., Sastry, N., & Tyson, G. (2021). *Exploring content moderation in the decentralized web: The Pleroma case.* https://arxiv.org/abs/2110.13500

269. Sokoto, S., Balduf, L., Trautwein, D., Wei, Y., Tyson, G., Castro, I., Ascigil, O., Pavlou, G., Korczyński, M., & Król, M. (2024). Guardians of the galaxy: Content moderation in the interplanetary file system. *Proceedings of the 33rd USENIX Security Symposium.* https://www.usenix.org/system/files/usenixsecurity24-sokoto.pdf

270. Thiel, D., & DiResta, R. *Child safety on federated social media.* Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:vb515nd6874/20230724-fediverse-csam-report.pdf

271. Bin Zia, H., Raman, A., Castro, I., Anaobi, I., De Cristofaro, E., Sastry, N., & Tyson, G. (2022). *Toxicity in the decentralized web and the potential for model sharing.* https://arxiv.org/abs/2204.12709

272. Hassan, A., Raman, A., Castro, I., Bin Zia, H., De Cristofaro, E., Sastry, N., & Tyson, G. (2021). *Exploring content moderation in the decentralized web: The Pleroma case.* https://arxiv.org/abs/2110.13500

273. Lou, X., & Hwang, K. (2009). Collusive piracy prevention in P2P content delivery networks. *IEEE Transactions on Computers, 58* (7). https://ieeexplore.ieee.org/document/4775892

274. Anaobi, I., Raman, A., Castro, I., Bin Zia, H., Ibosiola, D., & Tyson, G. (2023, February 12). *Will admins Cope? Decentralized moderation in the fediverse.* https://arxiv.org/abs/2302.05915

275. Sokoto, S., Balduf, L., Trautwein, D., Wei, Y., Tyson, G., Castro, I., Ascigil, O., Pavlou, G., Korczyński, M., & Król, M. (2024). Guardians of the galaxy: Content moderation in the interplanetary file system. *Proceedings of the 33rd USENIX Security Symposium.* https://www.usenix.org/system/files/usenixsecurity24-sokoto.pdf

276. Liberatore, M., Levin, B., & Shields, C. (2010). *Strengthening forensic investigations of child pornography on P2P networks.* https://web.cs.umass.edu/publication/docs/2010/UM-CS-2010-043.pdf

277. Hurley, R., Prusty, S., Soroush, H., Walls, R., Albrecht, J., Cecchet, E., Levine, B., Liberatore, M., Lynn, B., & Wolak, J. (2013). *Measurement and analysis of child pornography trafficking on P2P networks.* https://web.cs.umass.edu/publication/docs/2013/UM-CS-2013-007.pdf

278. Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2016). iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digital Investigation, 18,* 50-64. https://www.sciencedirect.com/science/article/pii/S1742287616300779

279. Lou, X., & Hwang, K. (2009). Collusive piracy prevention in P2P content delivery networks. *IEEE Transactions on Computers, 58*(7). https://ieeexplore.ieee.org/document/4775892

280. Bovenzi, G. (2024). Content moderation in (decentralized) metaverses. *Proceedings of the International Congress Towards a Responsible Development of the Metaverse*, Alicante, June 13–14, 2024. https://catedrametaverso.ua.es/wp-content/uploads/2024/07/Content-moderation-in-decentralised-metaverses-BOVENZI.pdf

281. eSafety Commissioner. (2021). *Decenstralisation—Tech trends position statement.* Australian Government. https://www.esafety.gov.au/sites/default/files/2021-08/Decentralisation%20Position%20statement_1.pdf?v=1728351391299

282. IBM. (n.d.). *What is quantum computing?* https://www.ibm.com/topics/quantum-computing

283. Zander, J. (2024). *Advancing science: Microsot and Quantinuum demonstrate the most reliable logical qubits on record with an error rate 800x better than physical qubits.* Official Microsoft Blog. https://blogs.microsoft.com/blog/2024/04/03/advancing-science-microsoft-and-quantinuum-demonstrate-the-most-reliable-logical-qubits-on-record-with-an-error-rate-800x-better-than-physical-qubits

284. Chen, S. (2024, September 11). Google says it's made a quantum computing breakthrough that reduces errors. *MIT Technology Review.* https://www.technologyreview.com/2024/09/11/1103828/google-says-its-made-a-quantum-computing-breakthrough-that-reduces-errors/?utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_term=&utm_content=09-18-2024&mc_cid=9ec6b000b1&mc_eid=6c8f75ac82

285. DWave. (n.d.). *Quantum in life sciences.* https://www.dwavesys.com/solutions-and-products/life-sciences

286. Bova, F., Goldfarb, A., & Melko, R. (2021). Commercial applications of quantum computing. *EPJ Quantum Technology, 8,* article 2. https://epjquantumtechnology.springeropen.com/articles/10.1140/epjqt/s40507-021-00091-1

287. KETS. (n.d.). *Quantum key distribution.* https://kets-quantum.com/quantum-key-distribution

288. eSafety Commissioner. (n.d.). *eSafety Strategy 2022-25.* https://www.esafety.gov.au/about-us/who-we-are/strategy

289. National Institute of Standards and Technology. (2024, August 13). *NIST releases first 3 finalized post-quantum encryption standards.* https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards

THORN

thorn.org | info@thorn.org

weprotect Global Alliance

weprotect.org | info@weprotecga.org