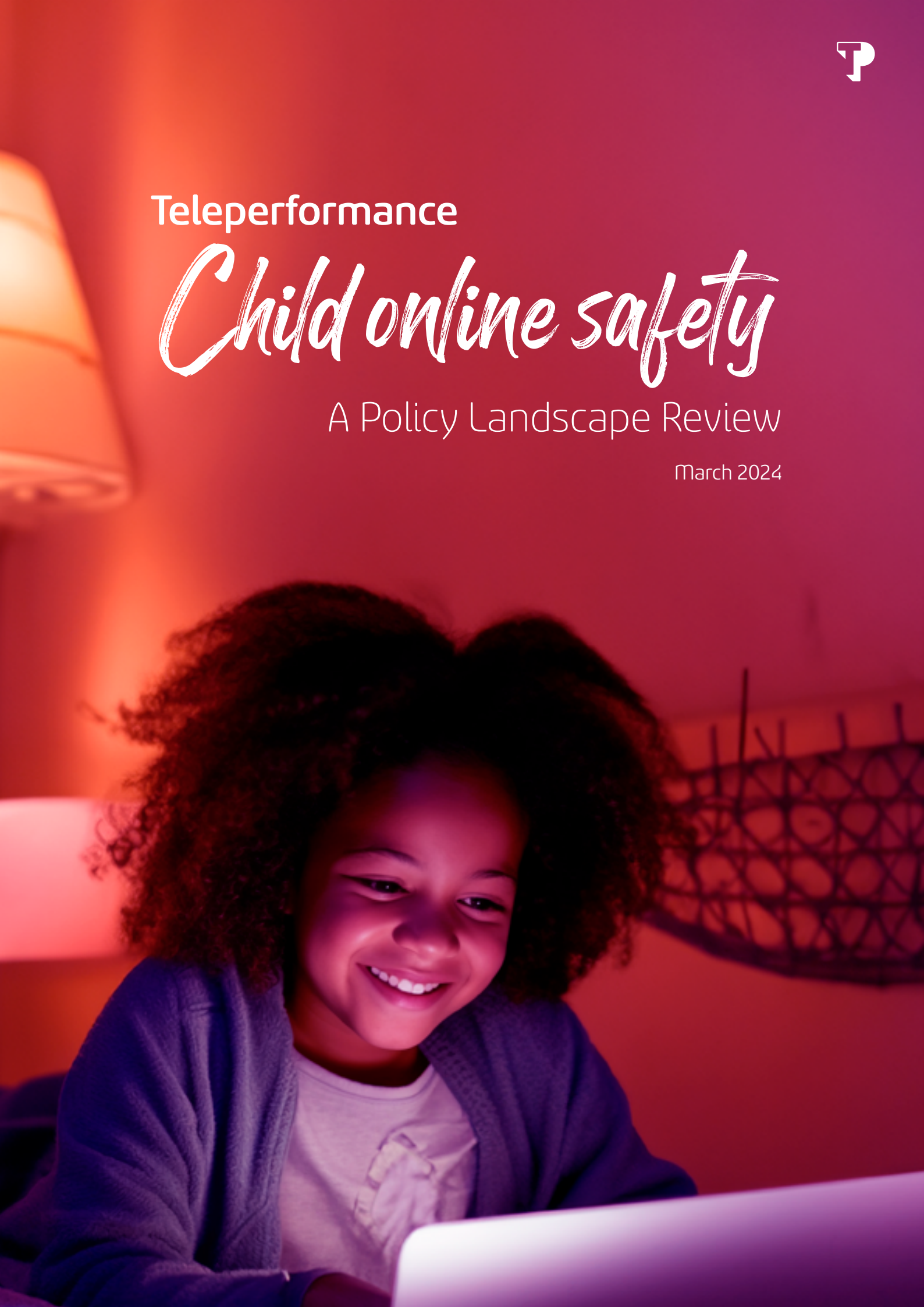
Teleperformance

Child online safety

A Policy Landscape Review

March 2024



Introduction and Context

In July 2022, Teleperformance launched its first-ever Trust & Safety Advisory Council to bring together external experts in the online safety space and to drive collaboration in areas that are of critical importance to Trust & Safety. One of these areas is child online safety. This report aims to provide an overview of the key components of the child safety landscape as outlined by members of this Council.

EXECUTIVE SUMMARY

The child online safety landscape is growing in complexity and importance. With an evolving threat landscape facing children on the web, new regulation emerging, and a multitude of tech solutions being offered, it can be difficult to navigate the core challenges and opportunities to improve child safety.

Moreover, there is currently misunderstanding about the harms that children face online and what solutions will actually be beneficial, considering an evidence-based approach. Platforms have differing policies around minor safety, and governments around the world are proposing a variety of regulations.

The goal of this landscape review is to highlight some of the trends, challenges, and opportunities for child online safety by reviewing existing initiatives, policies, and practices in this area. This landscape review covers the following topics:

Section 1: Harms to Children

This section focuses on providing an overview of how harms to children are evolving using the 4C framework as the basis for categorization across:

- Content – Child as Recipient (e.g., exposure to eating disorder content)
- Contact – Child as Participant (e.g., grooming)
- Conduct – Child as Actor (e.g., bullying)
- Contract – Child as Consumer (e.g., sextortion)

It discusses new trends in each area (e.g., AI-generated harms) and provides examples of how these harms manifest on various types of platforms. The focus of this paper will be on harms related to sexual exploitation/abuse and other social and psychological harms, including cyberbullying, self-harm, and exposure to inappropriate (e.g., violent) content.

Section 2: Current Regulations and Legislative Proposals

This section highlights the most significant child safety-related regulations and legislative proposals, spanning areas such as age verification, platform bans, ad targeting toward children, age restrictions, privacy, and overarching regulations such as the Digital Services Act and the Online Safety Bill in the UK covering areas spanning content moderation, risk assessments, and transparency reporting, which impact children's online experiences.

Section 3: Platform Policies and Initiatives

This section provides an overview of key child safety policies, highlighting commonalities and differences in areas that are core to child online safety (e.g., CSAM, cyberbullying). It also highlights the role of additional initiatives/ features beyond platform policies that impact kids' safety, such as parental controls, insights generated from Youth Councils, and the overarching processes for reporting and moderating to enforce child safety.

Section 4: Multistakeholder Initiatives and Efforts in the Child Safety Space

This section briefly describes the mandate of the various industry consortiums, non-profits, hotlines, and other relevant initiatives in the child safety landscape and organizes them according to their function and mandate, including but not limited to organizations such as the WeProtect Global Alliance, INHOPE, Thorn, the Tech Coalition, the International Centre for Missing and Exploited Children, and the National Center for Missing and Exploited Children.

Section 5: Challenges to Overcome and Solutions to Explore

This section will highlight some of the historical and current challenges to advancing child safety further, touching on issues of regulatory fragmentation, trade-offs between privacy and safety, and age determination. It will also highlight potential solution areas that could drive accelerated, positive impact in this space, from new technology, research areas, adoption of safety by design principles, and content moderation enhancements.

TABLE OF CONTENTS:

SECTION 1: HARMS TO CHILDREN

No paper on children's online safety can provide a complete picture if it is focused only on the risks they face. Indeed, research from UNICEF highlights that both opportunities and risks to children go hand in hand. The report shows that "children who participate in more online activities tend to have better digital skills compared to those who engage in fewer activities. But the results also show that children who participate in more online activities tend to experience more risks as a result."

For those whose work is addressing the risk side, however, it is important to understand that end of the child's online wellbeing spectrum. So let's consider the "4 Cs" risk framework first developed by EU Kids Online and later augmented by CORE, the global research network that grew out of EUKO's work:

- Content: The child as the recipient, as in seeing, reading, and viewing inappropriate material such as pornography, graphic violence, self-harm, disordered eating, gambling, etc.
- Contact: The child as the participant, as in sharing, creating, contributing, and participating in content and behavior that is inappropriate or illegal, from sharing nudes to grooming
- Conduct: The child as the actor, as in perpetrating, inciting, and otherwise engaging in harmful conduct such as harassment or cyberbullying
- Contract: The child as the consumer, as in being subjected to data or identity theft, extortion, and other commercial or transactional harms from fraudulent scams to sextortion

	Content Child as Recipient	Contact Child as Participant	Conduct Child as Actor	Contract Child as Consumer
Aggressive	Violent, gory, graphic, racist, hateful, and extremist content	Bullying, hateful or hostile peer activity, e.g., trolling, exclusion, shaming	Bullying, hateful or hostile peer activity, e.g., trolling, exclusion, shaming	Identify theft, fraud, phishing, scams, gambling, blackmail, security risks
Sexual	Pornography (legal and illegal), sexualization of culture, body image norms	Sexual harassment, sexual grooming, generation and sharing of child sexual abuse material	Sexual harassment, non-consensual sexual messages, sexual pressures	Sextortion, trafficking for purposes of sexual exploitation, streaming child sexual abuse
Values	Age-inappropriate user-generated or marketing content, mis/disinformation	Ideological persuasion, radicalization and extremist recruitment	Potentially harmful user communities, e.g., self-harm, anti-vaccine, peer pressures	Information filtering, profiling bias, polarization, persuasive design
Cross-Cutting	Privacy and data protection abuses, physical and mental health risks, forms of discrimination			

Source: https://core-evidence.eu/posts/4-cs-of-online-risk

Each of the Cs has sub-categories: "Aggressive" content and behavior (violence, extremism, hate, stalking, blackmail, identity theft, etc.); "Sexual" (legal and illegal pornography, sexual harassment, body image norms, non-consensual intimate image-sharing, sex-trafficking, etc.); "Values" (age-inappropriate user- or advertiser-generated content, radicalization and recruitment, mis/disinformation, profiling, manipulative design, etc.); and material that is "Cross-Cutting" across all the Cs (abuse of data and privacy, discrimination, psychological and physical safety, etc.).

While businesses often organize them into a more basic framework of illegal vs. legal (often referred to as "lawful but awful") harms, CORE gets more granular in a way that can be helpful for content policy and moderation. So, let's zoom in on the three Cs that represent physical and psychological harm: Content, Contact and Conduct. With all of them, it's important to remember that all youth are not equally at risk online and that those most at risk online tend to be the young people most at risk in offline life, for example, those marginalized at school and those experiencing mental health challenges, especially during the youth mental health crisis being reported in many countries.

Contact

Below is a perspective on what platforms are doing about online child sexual exploitation, but first, let's take a measure of the problem itself. The University of New Hampshire's Crimes Against Children Research Center this year published extensive research into the forms of CSE online, their prevalence and their impacts. The researchers found that 16% of young Americans have experienced at least one type of sexual abuse online before they turn 18. In that 16% of case offenders were under 18. "Peers made up a majority of offenders, and their impact (on victims) was just as great as adult offenders," according to the authors of the study published in the Journal of the American Medical Association Network Open.

The authors found that 13-17-year-olds are the most vulnerable age group, with girls the more frequent victims across the board (73%). Those perpetrators are "predominantly dating partners, acquaintances, and friends," they reported, with the most frequent type of victimization being sexual solicitation: unwanted sexual questions (18.8%), talk (16.9%), and requests (14.3%). Other prevalences among the 18 types and categories described by the CCRC include nonconsensual sharing of images, whether or not obtained consensually (7.2%); online grooming by adults (5.4%); sextortion (3.5%); and 3.1% for "revenge pornography," or non-consensual sharing of "sexts" with the intent to embarrass or humiliate the victim (not including showing off).

The CCRC's most recent study on this looked at what abuse situations have the greatest emotional impact on victims. The authors found that the involvement of (sexual) images, in particular, causes much distress, especially "non-consensual sharing, threatened sharing/sextortion, and non-consensual taking. That third type refers to sexual images taken when the victim was unconscious, intoxicated, distracted, or otherwise unable to consent, including manufactured ones called "deepfakes." The negative impact on victims is no less when the abuse comes from peers rather than adult offenders or when it comes from people the victims know rather than strangers, the CCRC reports. The authors also found that the degree of impact was not greater with more explicit sexual images. They wrote that the reason may be because the loss of control over the image or the sense of betrayal is more likely the source of distress than what is depicted in the image.

Conduct

The term "conduct" represents the fact that, in this category of online harm, people, based on how they conduct themselves online, are seen as perpetrators as much as victims. This is the "awful but lawful" category of harm. "Awful" for two reasons: because it is much more commonly experienced by minors than the criminal and illegal types of harm above and because it is traumatizing for them when peers bully, humiliate, harass, or marginalize them.

The US's Cyberbullying Research Center defines "cyberbullying" as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices." Minors who experience it are at increased risk for depression, anxiety, loneliness, and suicidal ideation, and studies on both sides of the Atlantic have shown a correlation between cyberbullying and self-harm, eating disorders, and substance abuse. Victims are also at increased risk for depression, anxiety, loneliness, and suicidal ideation.

According to research by the Cyberbullying Research Center across 11 national surveys over at least as many years, about 27% of US teens have been cyberbullied at some point in their lifetimes, with girls just as if not more likely than boys to experience it as both offenders and victims. In the UK and Europe, estimates range from 10% to 30% of teens experiencing this harm.

As for what this type of content looks like, according to the Center, the most common forms of cyberbullying that have taken place in the US this past year were expressed as:

- Someone posted mean or hurtful comments about me online (77.5%)
- Someone spread rumors about me online (70.4%)
- Someone embarrassed or humiliated me online (69.1%)
- Someone intentionally excluded me from a group text or group chat (66.4%)
- Someone repeatedly contacted me via text or online after I told them to stop (55.5%)

Generative AI

Of course, the latest technology to challenge the work of online harm mitigation is generative AI. As of this writing, it's very much "early days" for a full understanding of the harms it could bring. Besides the somewhat abstract existential risks that some pioneering AI engineers such as British-Canadian computer scientist Geoffrey Hinton have called out and fears of election-related disinformation much in the news, the harm that most concerned child protection advocates and policymakers in a number of countries were discussing most toward the end of 2023 concerned child sexual exploitation (CSE): synthetic images popularly known as "deepfakes."

So far, three forms of this AI-generated child sexual abuse material (AIG CSAM) have been identified. Two are nude or sexually explicit images and videos that are either a blend of AI-generated imagery and real minors' faces or other body parts or entirely AI-generated images depicting minors. Those first two forms are created by criminals or other bad actors for the purpose of extortion, grooming, or trafficking in child abuse imagery. A third is real or partly real images of minors created by peers abusing AI for that purpose in the form of harassment, bullying, or extortion. We don't yet know if entirely simulated child sexual abuse images, whether simulating photography or animated type material, will be deemed illegal by laws around the world.

The Internet Watch Foundation, a UK-based international child protection organization, reported that 7 out of 29 reports of suspected AIG CSAM it received during a five-week period in 2023 were confirmed as such. Thorn, a US-based anti-CSE organization, has conducted several studies related to deepfakes and CSE and reports that, though the proportion of AIG CSAM vs. the "traditional" human-generated type is still low, the numbers are growing fast. A June 2023 study by researchers at Thorn and Stanford University reported that "based on an internal study by Thorn, less than 1% of CSAM files shared in a sample of communities dedicated to child sexual abuse are photorealistic CG-CSAM (or AI-generated CSAM), but this has increased consistently since August 2022." The problem for content moderators and law enforcement is that the AI-generated imagery is getting increasingly sophisticated. "Thorn finds that approximately 66% (of the images found in those CSAM-trafficking communities) are highly photorealistic, but can currently be visually distinguished as being generated (e.g., due to roughness of skin edges or highly pixelated areas)," the report continued. There is no telling as yet when the arms race between deepfake creation and deepfake detection will end.

Emerging challenge: Social GAI

If it's early days for AIG CSAM, even more nascent is our understanding of what might be termed "social GAI." GAI is getting more and more social, or interactive. It's rapidly moving beyond the 1:1, human-plus-machine creation of content to 1:1 human+chatbot interaction and 1:many chat with groups of people bringing chatbots into their digital discussions. One example is people interacting with celebrity-based chatbots in social media and messaging applications. Another is chatbots in mental health apps, 10 of which were reviewed in a 2023 study published in the journal JMIR mHealth and uHealth. There are all kinds of potential for entertainment, learning, and mental health support in this interaction, but there are risks as well, ranging from misinformation to emotional dependency or harm to the inability of chatbots to detect signs of mental health crisis. Readers may remember when, in early 2023, a New York Times story of an early version of ChatGPT suggesting that maybe reporter Kevin Roose should leave his wife went viral. Though it doesn't seem to have affected Roose's mental health or marriage, more research is needed into the effects of potentially manipulative chatbots on children and vulnerable adults.

Economic and regulatory factors

There are two trends in online child protection that are not new but increasingly impactful moving into 2024. On the regulatory front, all eyes are on Europe's Digital Services Act and the UK's Online Safety Act, with increasing regulatory activity at both state and federal levels in the United States. Meanwhile, regulatory activity at an international level has grown since the 2022 launch of the Global Online Safety Regulators Network. This requires an increased focus from platforms to not only mitigate user risks, but also legal risks, to the extent that those do not completely overlap. Adding to regulatory pressures is the economic factor. The rise of generative AI seems to have ushered in a new "move fast and break things" phase, with investment both internal and external to businesses which leads to an increased focus on innovation. It is important that safety does not become a competing priority. One sign was the layoffs of tens of thousands of tech workers seen to have increased platforms' dependency on algorithmic moderation and heightened triaging for illegal harms at a time of increasing regulation. All of this is likely to lead to renewed and increasing calls for safety by design.

SECTION 2: CURRENT REGULATIONS AND LEGISLATIVE PROPOSALS

The global regulatory trend for children's online safety indicates a concerted effort to balance technological innovation with the safeguarding of children in the digital age, with a few themes consistently emerging in different markets and contexts.

Age verification

Age verification is the process or system designed to confirm the age of an individual, typically in relation to access to certain content, services, or products that may have age restrictions. This verification is commonly used in various online platforms, websites, or applications to ensure that users meet the minimum age requirements for specific activities or content consumption. Age verification methods can include providing official identification documents, entering a date of birth, using biometric data, or employing other technological solutions to confirm that a user is of a certain age. Age verification is often implemented to comply with legal regulations, especially in areas where there are age restrictions on certain types of content or services, such as gambling, alcohol consumption, or access to explicit content.

There are two main pieces of legislation in Europe driving this discussion: the Digital Majority Law in France and the Online Safety Bill in the UK.

France approved a new law in July 2023, the Digital Majority Law, that requires social media platforms to verify users' ages and obtain parental consent for those under 15 years to protect children online. The providers must decline to register minors under 15 on their services, except with the express consent of one of the minor's legal guardians. The law will also allow parents to request suspension of accounts belonging to their children under 15 and require sites to offer tools to limit the time children spend on the platform.

To do that, providers must activate a device to monitor the time of use of their service by the minor upon registration, and the user must be regularly informed of his/her time of use via notifications.

This legislation is part of a string of recent moves by the French government to reduce children's screen time and protect them from cyberbullying, pornography, cyberstalking, unattainable beauty standards, and the attention-grabbing addictive nature of the platforms. Arcom (the French telecom regulator) is responsible for drawing up guidelines and detailing the technical procedures.

However, in a report on online age verification, the French Data Protection Authority -Commission Nationale de l'Informatique et des Libertés (CNIL)- analyzed several existing solutions for online age verification and highlighted challenges to potentially meeting data privacy and security requirements. While some companies providing solutions in this area disagree, there is ongoing discussion on this front.

The UK introduced its Online Safety Bill in October 2023, requiring companies to keep children away from harmful content by enforcing aging limits and age-checking measures (age assurance), either by asking for government-issued documents (age verification) or using biometric data, such as face scans to estimate their age (age estimation). Self-declaration will not be accepted for compliance purposes.

Compliance will be compulsory unless the platform's terms of service explicitly prohibit the content being addressed. Providers can even be required to distinguish between children of different ages to determine whether they can be permitted to access certain content. Ofcom (UK telecoms regulator) is in charge of detailing and enforcing the rules.

Cell phone and platform bans

According to a UNESCO 2023 report, one in four countries has implemented laws or policies restricting or outright banning student cell phone use. The prevalence of such bans is notably higher in Central and Southern Asia. Various nations, including Latvia, Mexico, Portugal, Spain, Switzerland, the US (with a recent statewide ban in Florida), Canada (in Ontario), and the UK, have enforced either full or partial bans on mobile phones in educational settings. The report emphasizes the contentious nature of this issue, highlighting concerns over data privacy, safety, and well-being.

Beyond cell phone bans, some schools in the United States are prohibiting social media platforms, particularly TikTok - a concern also outlined in an executive order from March 2023.

Source: Profiles enhancing education reviews (PEER)

One in seven countries ban the use of mobile phones in schools by law

Percentage of the countries taking measures to ban mobile phones in schools, by tool, 2022

Restrictions to targeted ads to children

According to COPPA and the DSA, placement-based advertising on children-focused videos is not forbidden; however, targeting children based on user profiling is restricted (in the case of COPPA) and forbidden (in the case of DSA). In Brazil, the so-called Fake News Bill also restricts targeted advertising for children.

The DSA prohibits online platforms from showing targeted advertisements to minors based on the use of their personal data. "Providers of online platforms shall not present advertisements on their interface based on profiling as defined in Article 4, point (4), of Regulation (EU) 2016/679 using personal data of the recipient of the service when they are aware with reasonable certainty that the recipient of the service is a minor." The DSA also puts an obligation on online platforms to assess systemic risks that might harm children's rights and mental well-being and health. In the UK, new rules to protect children online demand social media platforms, websites, and services like advertising display networks to take tougher action to stop children from seeing age-restricted adverts for products like alcohol or gambling.

In the US, COPPA (Children's Online Privacy Protection Rule) restrains advertisers and content owners from collecting any personal information (which includes cookies and other persistent identifiers) from children under 13 years of age without verifiable parental consent. Recently, President Joe Biden called for ban of online ads targeting children altogether (July 2023). Ed Markey (D-Mass.) plans to reintroduce the Children and Teens' Online Privacy Protection Act this Congress (COPPA 2.0).

In Utah, the Social Media Regulation Act from October 2023 states that starting on March 1, 2024, social media companies cannot: cannot (1) collect a minor's data, (2) target a minor's social media accounts for advertising, and (3) target minor's social media accounts with addictive designs or features. Social media companies must:

- Verify the age of a Utah user seeking to maintain or open a social media account
- Get the consent of a parent or guardian for Utah users under the age of 18
- Allow parents or guardians full access to their child's account
- Create a default curfew setting that blocks overnight access to minor accounts (10:30 pm to 6:30 am), which parents can adjust
- Protect minor accounts from unapproved direct messaging
- Block minor accounts from search results

NetChoice has filed a lawsuit to block this law from going into effect, alleging that the law violates the First and Fourteenth Amendments by limiting residents' "right to share and receive information online on their willingness to hand over their most sensitive, personal data to third-party age-verification companies."

A new federal bill, Protecting Kids on Social Media Act, aims to establish a national minimum age for social media use and require tech companies to get parents' consent before creating accounts for teens, reflecting a growing trend at all levels of government to restrict how platforms engage with young users.

- This bill requires social media platforms to verify the age of account holders and limits access to such platforms by children. Specifically, social media platforms (1) must verify the age of account holders, (2) may not allow an individual to create or continue to use an account unless the individual's age has been verified, and (3) must limit access to the platform for children under the age of 13. Social media platforms may not use or retain any information collected during the age verification process for any other purpose.
- Further, platforms must take reasonable steps to (1) require affirmative consent from the parent or guardian of a minor who is at least 13 years old to create an account for the minor on the platform and (2) provide the parent or guardian with the ability to revoke such consent.
- Social media platforms may not use an individual's personal data in an algorithmic recommendation system unless the individual is at least 18 years old according to the platform's age verification process.

Most recently in the US, New York has sued social media companies "alleging their platforms' designs exploit young users' mental health and cost the city $100 million in related health programs and services each year." Whether this suit will be successful, remains to be seen.

Children's safety by design:

Children's safety by design is another hot topic related to privacy concerns.

The UK's Age-Appropriate Design Code takes account of the standards and principles set out in the United Nations Convention on the Rights of the Child (UNCRC), and sets out specific protections for children's personal data in compliance with the provisions of the GDPR.

Some of the standards are: best interest of the child; data protection impact assessments; age appropriate application; Transparency; children's personal data use is prohibited in ways that have been shown to be detrimental to their well-being; policies and community standards; High privacy as default settings; data minimization; disclosure of children's data is prohibited unless you can demonstrate a compelling reason to do so; switch geolocation options off by default; parental controls and information about them; switch options that use profiling off by default; nudge techniques to lead or encourage children to provide unnecessary personal data or weaken or turn off their privacy protections are prohibited; connected toy or device must include effective tools to enable conformance; prominent and accessible online tools to help children exercise their data protection rights and report concerns.

Similarly, the California Age-Appropriate Design Code Act, requires online businesses likely to be accessed by children - defined as any user under the age of 18 as determined through "age assurance" methods - to default privacy settings to the highest level and complete an impact assessment before any new products or features are made publicly available. It is important to note that a US Court has blocked this law, stating that it probably violates the First Amendment.

As concerns arise regarding the effects of social media usage on children's mental health, other state legislators are introducing measures to protect children while using the internet and internet-based forms of communication, including social media. The legislation includes bills and resolutions that (1) create study commissions and task forces, (2) require age verification or parental consent to open social media accounts, (3) add digital and media literacy courses or curriculum for K-12 students, (4) require age verification to prevent children from accessing explicit or harmful materials from websites.

As of 2023, 35 American states and Puerto Rico have pending legislation, and 11 states have enacted bills or adopted resolutions.

Across the U.S., the Kids Online Safety Act (KOSA) has gotten significant bipartisan support in the Senate; the most interesting parts of this bill are requirements around duty of care, safety by design, and default settings. A helpful deep dive into this bill can be found here.

Overarching regulations' impact on children

The DSA gives clear recognition of the rights of the child, as of:

- The swift removal of illegal online content, which includes child sexual abuse material, illegal hate speech, terrorist content, or an illegal product. Victims of online harassment will be better protected against unlawful non-consensual sharing of private images with immediate takedowns
- Risk assessments for impacts on rights, including those of children.
- A ban on targeted advertising at children and restrictions on data harvesting for profiling
- Greater understandability. For services that are primarily directed at minors or are predominantly used by them, the platform must explain terms of use in a way that minors can understand.

- Online platforms may put in place age verification measures, as discussed earlier, to control who can accesas discussed earlier, to control who can access their services, parental controls so that parents and guardians can help protect their children against the risk of exposure to harmful content, and tools where users can signal abuse or get support

Similarly, the UK Online Safety Bill puts particular emphasis on protecting children online, with additional requirements to prevent children from accessing "harmful and age-inappropriate" content. Online platforms must prevent children of any age from encountering "primary priority content harmful to children" and to "protect children in age groups judged to be at risk of harm from other content that is harmful to children (or from a particular kind of such content) from encountering it by means of the service"

The primary priority pieces of content are pornography, content encouraging, promoting, or providing instructions for suicide, self-harm (including poisoning), and eating disorders, which prohibit children from accessing them. Other types of priority content are anything depicting violence against people or animals (including fictional animals), bullying, abusive content related to several protected characteristics, content that promotes dangerous stunts (such as the "challenges"), and content that encourages people to "ingest, inject, inhale or in any other way self-administer" a physically harmful substance, or any substance in quantities, which would be harmful, health and vaccine misinformation, harassment, and violence (this content must be "age-appropriate" for children).

Platforms also now have to consider how to protect children from "features, functionalities, or behaviors enabled or created by the design or operation of the service." Platforms will also have to conduct a risk assessment to explain how they will address children of any age and those in age groups judged to be at risk of harm. There is also a transparency reporting obligation (once a year by Ofcom request).

In Australia, the Basic Online Safety Expectations (known as 'the Expectations') outlines the AU government's expectations for social media services, relevant electronic services, and designated internet services. Some of the Expectations for providers include:

- Protecting children from content that is not age-appropriate, like pornography
- Making sure the default privacy and safety settings of services targeted at, or used by children, are robust and set to the most restrictive level.

An ongoing amendment process, which closed in February 2024, asked for inputs on these expectations. One of the key themes was: "The best interests of the child and restricting access to age-inappropriate materials online."

In Brazil, the overarching Fake News Bill also states that providers must have "the best interest of the child" as a parameter of their services and adopt measures to ensure a high level of privacy, data protection, and security for children.

The next frontier is artificial intelligence (AI). The EU AI Act agreed text sets out due diligence requirements for high-risk AI applications, including as regards respect for children's rights based on UNCRC General comment No. 25.

SECTION 3: PLATFORM POLICIES AND INITIATIVES

Platforms have long maintained and worked to enforce policies against certain forms of problematic behavior by users. While there exists some variation given UX, UI, engineering, extant community norms, and the specific way in which interactions take place, some policy consistencies can be found across platforms, given their shared commitment to building spaces where youth can participate in safe, healthy, and positive ways.

To begin, certain forms of content are universally prohibited. Platforms generally disallow any content that endangers or exploits children. Typically, warnings are not issued when this content is posted or shared. Instead, computational methods in the form of hash mashing and AI are used to proactively identify CSAM in the form of photos and videos, and domestic or international law enforcement is immediately summoned to assist and respond. Given the significantly harmful and illegal nature of such content, as well as the extreme vulnerability of children around the world, platforms must remain exceedingly vigilant in their efforts to combat CSAM and protect this population of potential victims.

Harassment, cyberbullying, intimidation, and related forms of interpersonal abuse are also barred across major platforms in the interest of creating and preserving safe and respectful online communities for young people. TikTok, Snapchat, Twitch, Discord, YouTube, and many other mainstream platforms all explicitly prohibit the most direct forms of mistreatment of other users. Relatedly, the targeted harassment of another individual or group based on identity-based attributes such as gender, sexual orientation, identity or expression, race, religion, nationality, or disability is often characterized as hate speech and consequently condemned and disallowed. Platforms generally also formally forbid other problem behaviors such as attempts at fraud and deception, doxing, swatting, or IRL (in real life) harm, impersonation, content that depicts or glorifies self-harm or self-injury, and the unauthorized sharing of intimate images.

Platforms have also offered specialized subsets of controls to help parents and guardians support the youth under their care as they explore and embrace online environments. By way of example, TikTok provides a feature called "Family Pairing" that allows for a child and parent account to be linked, thereby giving parents additional safety controls to manage screen time, prevent mature content in their feeds, and stay apprised who they are following and who is following them. Similarly, Snapchat offers a Family Center that affords much of the same after parent and child accounts are formally tethered together. Twitch provides a suite of reporting, blocking, and moderation features that can help both streamers and viewers safeguard their experiences, along with a specific guide for parents and educators to enhance their ability to support youth. Discord has joined its peers to add functionality that allows parents to know which communities their teens belong to, and which users their teens are interacting with.

YouTube allows for a "supervised" account so that parents and guardians can select content settings to limit the videos and music their child can access, restrict the features that can be used, modify the ads they might see, and change the default settings on the account. It should be made clear, though, that parental controls across all of these platforms are limited by a child's prescribed right to privacy (from the perspective of the platforms). That is, parents are not given access to the private messages sent and received by youth, provided with the contents of comments or posts their child exchanges, or given any additional information to identify the users with whom their teen is interacting.

Furthermore, many platforms have created and built out Safety Centers to provide resources and guidelines to help their users understand and navigate their apps and sites in safe and informed ways. These "help" portals available on the corporate website or within the app often include clear, screenshot-supplemented directions on how to report inappropriate content, block and mute other users, set up privacy controls, and implement other protections or guardrails. They also offer strategies, how-to videos, tips, and sometimes even experiential activities for parents, guardians, mental health specialists, health care workers, and other youth professionals to help teens learn how to use their platforms safely.

Finally, it is refreshing to see a number of platforms begin to formally create youth councils to solicit their unique perspectives and insights to shape policy, product development, and educational initiatives. Given that youth today are experts at knowing what it's like to be a youth today, such an approach is laden with promise. Listening to, validating, and encouraging youth voices not only meaningfully enhance the trust and safety efforts of these companies — whose viability is often inextricably intertwined with the widespread affection and adoption of their platform by young people — but also help develop the next generation of leaders in this field. The youth should have a seat at the proverbial table to inform relevant decision-making, and can help catalyze innovative approaches in the ways that platforms serve and support their users.

SECTION 4: MULTISTAKEHOLDER INITIATIVES AND EFFORTS

Multistakeholder initiatives in the child online safety space are numerous, which can be mind-boggling even for experts in the field. In this section, we try to list and organize them according to their main mission. We have identified five main roles:

- Initiatives advocating for stronger regulations and increased scrutiny and safety processes
- Helplines providing specific guidance to minors and parents
- Hotlines issuing takedown of Child Sexual Abuse Material available online (reported by the public or found proactively)
- Initiatives aiming at developing tools and technologies
- Funding initiatives

As most stakeholders are engaged in more than one role, this section cannot do justice to all activities of all stakeholders. At the risk of subjectivity, we are highlighting the initiatives that are considered the most active for the role for which they are best known.

1. Advocacy and Policy

The WeProtect Global Alliance can be described as a meta-organization advocating for the protection of children online. With more than 270 members, governments, industry and international NGOs (many having tens of thousands of members), it has the broadest membership of stakeholders and is a natural starting point to map this rich landscape. Some of its most influential members in advocacy - some of them can be described as activists - are Brave Movement, Canadian Centre for Child Protection, ECPAT International, Family Online Safety Institute, International Centre for Missing and Exploited Children, IJM, Lucy Faithfull Foundation, Marie Collins Foundation, Missing Children Europe, or Save the Children.

Industry-oriented organizations, such as the Technology Coalition, work alongside these efforts to drive company collaboration and action to tackle CSAM online.

2. Helplines and Awareness Nodes

Insafe, an EU initiative, coordinates 29 helplines from the EU (plus the UK and Norway) and is connected with 13 countries from the Balkans, LatAm (Brazil, Mexico), Middle-East. It is the largest connector of helplines. Outside of this network, NCMEC in the US is the largest helpline and a global reference.

3. Hotlines

INHOPE, also an EU initiative, is a network of 54 hotlines from 50 countries. It develops best practices for existing and new hotlines, a common ontology and operates a sharing platform of URLs and hashes of CSAM. The hotlines processing the highest volume of reports are IWF (UK), NCMEC (US), Offlimits (NL), and Point de Contact (FR). In addition, the Canadian Centre of Child Protection is a major stakeholder among hotlines.

4. Technology

In terms of operational sharing of CSAM, the best-known projects using advanced technology are the INTERPOL ICSE database (for victim identification purposes), NCMEC CyberTipline reports by Electronic Service Providers (the largest volume of reports by far, with close to 32 million reports in 2022), ICCAM database by INHOPE (which connects all hotlines for international sharing of reports), Project Arachnid by the Canadian Centre of Child Protection (it brings together 14 hotlines members of INHOPE and the Canadian Center to detect proactively CSAM online), and the Offlimits Hash-Check server (for automatic takedown of content by cloud providers in the Netherlands).

Some initiatives are focusing on developing technologies. They can have a broad mandate, such as Thorn, the most notable NGO in this category. Others can be more specialized in certain use cases: PROJECT VIC International is focused on technologies to rescue children from sexual exploitation, while Project AviaTor helps prioritize all aspects of NCMEC reports so that Law Enforcement Agencies (LEAs) can focus on identifying perpetrators and saving victims.

5. Funding initiatives

All these initiatives are in the constant quest for funding. Funding from government, tech companies, platforms, and individual donors (undisclosed) are the most common sources of funding. A type of funding that is still underdeveloped in the relatively recent area of child safety online is funding by foundations. A landmark initiative is Safe Online, formerly known as End Violence Against Children. It invested more than $75 million in 90 projects in 85 countries, all projects being transparently displayed on their website.

SECTION 5: CHALLENGES TO OVERCOME AND SOLUTION TO EXPLORE

Despite progress made in improving safety for children online through the efforts described above, there are a number of challenges to overcome.

Fragmentation

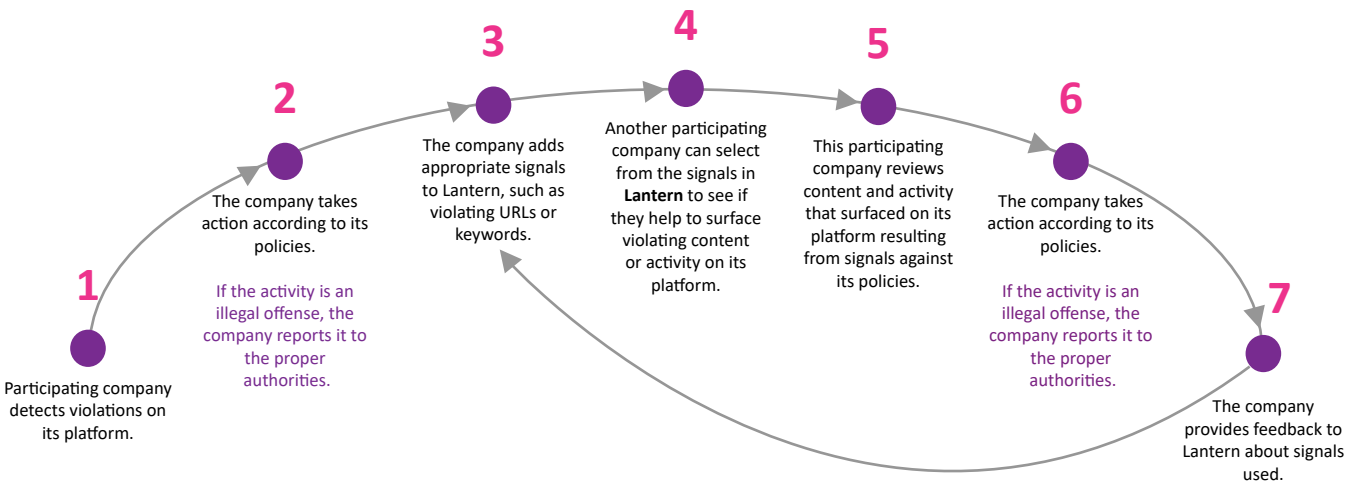
One of the biggest is the fragmentation and lack of cohesiveness in child safety regulation and even the definitions of CSAM across jurisdictions. This is a known problem and solutions are being developed on to address this. For example, INHOPE is driving the Global Standard project to bring together the skills and experience of global experts to develop a common categorization of CSAM. Anticipated to have a major impact on CSAM detection and removal globally, the project will create a common ontology of CSAM categorization, which will facilitate automated translation between and among different categorization schemas in use globally and across sectors.

Information Sharing

Another challenge is that beyond hash sharing for known CSAM through tools like PhotoDNA, there has historically been little other tangible collaboration across the industry to share information and get ahead of bad actors who are endangering children. One project aiming to change this is Project Lantern, which is creating a mechanism to share signals to better tackle online child sexual abuse and exploitation.

According to the Tech Coalition, "signals can be, for example, information tied to policy-violating accounts like email addresses, usernames, CSAM hashes, or keywords used to groom as well as buy and sell CSAM. Until now, no consistent procedure existed for companies to collaborate against predatory actors evading detection across services. Lantern fills this gap and shines a light on cross-platform attempts at online child sexual exploitation and abuse, helping to make the internet safer for kids. The program can enable the increase of prevention and detection capabilities; speed up the identification of threats; build situational awareness of new predatory tactics; and strengthen reporting to authorities of criminal offenses."

HOW IT WORKS:



Source: https://www.technologycoalition.org/newsroom/announcing-lantern

Age Assurance and Age Determination

As introduced earlier in this report, one of the most critical topics in the child safety policy space is both the need for and the appropriate solution for verifying the age of users to gain access to platforms. Age verification is a critical component in efforts to protect children, enhance online safety, and maintain ethical and legal standards in the digital era. Age verification is necessary to ensure compliance with laws and regulations related to explicit content, particularly when minors are involved. Enforcing age restrictions inhibits the access and distribution of online child sexual exploitation and abuse, contributing significantly to the overall safety of online spaces. Besides serving as a protective barrier, age verification measures make it more challenging for minors to access harmful content, acting as a deterrent against unintentional exposure and enhancing the protection of minors.

When it comes to age assurance or verification, recent regulations in the UK requiring enforcement of age gates is receiving mixed reactions. Some stakeholders highlight a potential negative impact on privacy, while others highlight this verification as a necessary and practical step to ensure that policies around the age of use are appropriately enforced. This is because the commonplace mechanism of simply requesting entry of a birthdate is easily bypassed. The Digital Trust & Safety Partnership (DTSP) has produced a set of guiding principles on Age Assurance highlighting the trade-offs between different options for age assurance. While some platforms have already started using age estimation technology to verify the age of users, others still rely on more basic measures such as entry of birth date. There isn't consensus on what constitutes an appropriate balance between effectively enforcing age restrictions and minimizing minor collection while creating a good user experience. This speaks to a broader tension between improving privacy (e.g., through end-to-end encryption) and child safety (e.g., inability to detect harmful or illegal content in encrypted channels), which is still unresolved.

Separate from age verification is the issue of age estimation. Age determination is a fundamental aspect of child protection that ensures children receive legal protection. A child's age establishes legal parameters and shapes regulatory frameworks. Accurate age determination is essential for protecting children from exploitation and plays a pivotal role in maintaining legal standards and prosecuting individuals involved in child exploitation and abuse.

This involves identifying whether an individual in a video or image is underage. This is particularly important in the case of sexual content where the difference between an adult (e.g., 18-year-old) and a child (e.g., 17-year-old) can determine the legality of the content (i.e., whether it can be considered CSAM). Currently, this is done through both manual estimation by people and/or by technology is an important discussion, especially with new research highlighting that AI tech exaggerates biases in facial age perception more than humans. While advancements in technology have improved accuracy of age estimation products over time, the majority of stakeholders understand the criticality of manual, human review for verification, given the potential implications of reporting such content to authorities. One of the main opportunities for improvement is driving industry alignment and best practices for improving human age estimation.

Prevention vs. Reactiveness

One of the challenges with advancing child safety is focusing efforts on prevention and addressing the root causes of "demand" for harmful material. John Tanagho, Director of International Justice Mission's Center to End Online Sexual

Exploitation of Children, shares his perspective from his written US Congressional testimony from September 2023:

"Research by Michael Salter (Associate Professor of Criminology at New South Wales University) reveals strong links between viewing violent and extreme adult pornography and child sexual abuse across the US,UK, and Australia. Professionals at the UK-based charity Lucy Faithful Foundation, who help men overcome viewing CSAM, report that a common pathway to offending is heavy pornography use, leading to habituation and desensitization. Analysis of qualitative data in a CSAM dark web survey by Finnish NGO Protect Children revealed that respondents may begin searching for CSAM when they are 'bored,' 'unexcited,' or 'tired' of adult pornography. In addition to achieving prevention through offender accountability leading to deterrence, governments should seek to go upstream to understand and disrupt some pathways to CSAM offending."

Research from the WeProtect Global Alliance's Global Threat Assessment 2023 also highlighted"emerging evidence of an association between the frequent viewing of pornography and progression to viewing child sexual abuse material" and that "escalation from legal content to child sexual abuse material can occur due to 'boredom' with legal content, increasing desensitization, or progression onto more extreme material to continually achieve sexual gratification."This highlights the complexity and intricacy of the challenge and one rationale for keeping adult content off mainstream social platforms; it also further emphasizes the need to proactively use technology to detect both existing and novel CSAM at scale and keep these off platforms.

Safety Guardrails - Industry Standards

A dynamic digital landscape and a lack of industry standards pose a multifaceted challenge for children online. Child protection extends beyond international borders, and variations in regulations across jurisdictions complicate the task of creating universally applicable standards that respect local laws. Overcoming these challenges demands international cooperation, regulatory initiatives, and partnerships involving the tech industry, law enforcement, and child protection organizations. In conjunction, developing industry standards that can keep pace with these changes is a constant challenge.

Efforts to combat online harms to children include the development of advanced content detection algorithms, international cooperation, stricter regulations, and mandatory reporting mechanisms. However, the evolving nature of technology means that addressing these challenges requires ongoing adaptation and innovation in law enforcement and child protection efforts. Nevertheless, establishing comprehensive and universally applicable industry standards that remain current for enhancing child safety remains a complex and continuous pursuit.

There are differing views on the safety guardrails that should set the minimum standards for "safety by design" for children. While regulation sets this baseline in many instances (e.g., high privacy setting by default for children users based on the UK's age-appropriate design code), in many cases, a multitude of important design choices for children's online experience is left to platforms. Working to have an industry standard for these choices, separate (and above and beyond) regulatory requirements, could be helpful in raising the bar for safety.

Technology advancements, while posing new risks, are also playing an important role in advancing solutions for improving online safety. Emerging technology has significantly impacted the proliferation of explicit content, presenting multifaceted challenges in creating a safe online environment for children. Easier access and distribution mechanisms facilitate quick sharing and wider dissemination of explicit material, specifically child sexual abuse material (CSAM).

Technological tools, including virtual private networks (VPNs), allow offenders to conceal their

identities and locations, impeding law enforcement efforts. The dark web and encrypted networks provide a clandestine space for the anonymous sharing and trading of child exploitation content, and private messaging apps with end-to-end encryption allow discreet communication among offenders. Advances in artificial intelligence (AI) contribute to the creation of convincing deepfake content, complicating age verification processes. The livestreaming of explicit content in real-time presents obstacles for law enforcement in identifying and intervening in cases of child exploitation and abuse. And the sheer volume of explicit material overwhelms identification and removal efforts.

Cryptocurrencies enable anonymous transactions, hindering financial traceability. The challenges stemming from emerging technology emphasize the urgent need for comprehensive measures to create a safe online environment for children.

On the flip side, improvements in precision and recall of AI classifiers are one of these top advancements. Not only does this improve the ability of platforms to remove violative content proactively, it also reduces the amount of content that requires manual review. Finding ways to improve content moderation processes and practices is a key component to advancing child safety, as described in this joint two-part report between ICMEC and Teleperformance.

With an ever-evolving landscape of threats toward children, child safety policies and their enforcement, initiatives, and regulation must take a comprehensive and evidence-based approach to enhancing child online safety.

Council Member Authors:

Anne Collier | Founder and Executive Director, The Net Safety Collaborative

Jean-Christophe Le Toquin | Managing Partner, SOCOGI

Robert (Bob) Cunningham | CEO, International Centre for Missing & Exploited Children

Sameer Hinduja | Co-Director, Cyberbullying Research Center, Professor, Florida Atlantic University

Natalia Paiva | Founder of Alandar Consulting, Former Head of Instagram Public Policy, LATAM, Meta

Council Member Acknowledgements:

Sarah T Roberts | Associate Professor of Information Studies, UCLA School of Information Studies

David Ryan Polgar | Founder, All Tech is Human

Ranjana Kumari | Founder and Director, Centre for Social Research, Chairperson of Women Power Connect

Noam Schwartz | CEO, ActiveFence

Dr. Fabro Steibel | Executive Director, Instituto de Tecnologia & Sociedade do Rio de Janeiro

Nighat Dad | Executive Director, Digital Rights Foundation

Teleperformance Authors:

Akash Pugalia | Global President, Gaming, Entertainment, Media, and Trust & Safety

Farah Lalani | Global VP, Head of Gaming, Trust & Safety Policy