Empowering and Protecting European Youth Online: Streamlining Legislation and Promoting Positive Digital Experiences

Dr. Sameer Hinduja and Farah Lalani

February 2025

Qualitative Research Partner

ThinkYoung

Suggested Citation

APA
Hinduja, S. & Lalani, F. (2025). Empowering and Protecting European Youth Online: Streamlining Legislation and Promoting Positive Digital Experiences. Available: [insert URL where you found this report here].

Keywords
online safety, social media, platforms, youth, children, teens, legislation, regulation, European Union, EU, Europe, UK, risks, harms, prevention, regulatory response, trust and safety, T&S

Photo Credits

Cover, Page 9, Page 37, Page 11: Freepik
Page 10, Page 22, Page 28, Page 43: Unsplash

Table of Contents

About the Authors

Dr. Sameer Hinduja is a Professor in the School of Criminology and Criminal Justice at Florida Atlantic University, Co-Director of the Cyberbullying Research Center, and Faculty Associate at the Berkman Klein Center at Harvard University. He is recognized for his research on the risks and harms of emerging technologies on populations of youth. Dr. Hinduja also trains technologists, educators, mental health professionals, and others on how to support youth well-being, promote positive behaviors, and build healthy communities online.

Farah Lalani is a Trust & Safety expert with over a decade of experience across technology, policy, and public-private cooperation, having previously launched the Global Coalition for Digital Safety at the World Economic Forum. She worked closely with regulators, policymakers and platforms to devise new principles, risk assessments, and policies to advance online safety. She was also previously the Vice President of T&S Policy at Teleperformance and a Fellow at the Internet Society. She was a co-author on many reports in the T&S space, including Safety by Design Principles to Tackle AI-Generated CSAM with All Tech is Human and Thorn. She serves as a working group member to develop an IEEE standard to tackle AI-generated harms to children.

Executive Summary

In recent years, there has been significant public discourse and political will to regulate children's access and interactions online. This has led to a wide range of new laws governing the use and access of technology, particularly with social media products for youth. This report highlights the context surrounding emerging social media regulation in Europe, assesses existing regulation, identifies the deficiencies and challenges with current approaches, and proposes a streamlined regulatory framework to improve online safety for youth. This framework was built based on a consideration of the multifaceted nature of online engagement, its ubiquity, and intersection with almost every area of life. It is based on the existing evidence about young people's experiences online, government concerns about the design and operation of social media products used by minors, and the constraints within which platforms creating these products must operate.

Research is clear that social media use can provide a broad set of benefits for young people and also pose a number of risks. Social media platforms facilitate social interaction and help youth meet their relational needs, provide spaces for knowledge acquisition and skill-building, allow for identity exploration and development (particularly for marginalized youth), and offer avenues for civic engagement and activism in ways that resonate with this population. However, these benefits coexist with a broad array of certain risks and harms. This includes inappropriate content (violence, sexual content, hate speech), cyberbullying and harassment, and child sexual exploitation and abuse. Moreover, emerging risks involving harassment in metaverse environments and the growing problem of AI-generated harmful content demand our attention and focused response. Many factors impact which youth are most at risk, but more integrated protective measures at both governmental and platform levels are needed to mitigate potential emotional, psychological, and behavioural impacts to all youth, particularly those who are most vulnerable.

Legislation in Europe focused on youth and social media continues to evolve, with the United Nations Convention on the Rights of the Child and the General Comment No. 25 on children's rights in relation to the digital environment serving as guidance for laws that impact children. Other initiatives that aim to provide guiding principles include the Better Internet for Kids (BIK+) strategy, the ePrivacy Directive, and the Age Appropriate Design Code from the UK's Information Commissioner's Office. Major regulation that has emerged in Europe include the Digital Services Act (DSA) which aims to prevent illegal and harmful activities online and the spread of disinformation across the European Union (EU). The Online Safety Act (OSA) in the United Kingdom (UK) puts a range of new duties on social media companies and search services to implement systems and processes to reduce risks their services are used for illegal activity, and to take down illegal content when it does appear. Finally, other legislation such as the General Data Protection Regulation (GDPR) and the Artificial Intelligence (AI) Act also seek to inform best practices surrounding major safety, security, and privacy issues that are critically important to this population. When assessing regulation across Europe comprehensively, this implementation presents significant challenges for platforms, which may hamper the effective safeguarding of youth because of fragmentation and overlapping requirements.

Given this backdrop, a new framework called 'SAFEST' (**S**afety, **A**utonomy and Choice, **F**ree Expression, **E**vidence-based Practices, **S**ecurity and Privacy, and **T**ransparency) has been developed as part of this effort to guide both regulator and platform action on youth online safety. It formulates priority actions for regulators to mandate across all platforms in order to promote well-being among this vulnerable population. Regulators must mandate industry-wide standards across key areas that are aligned with the SAFEST framework.

Importantly, regulators should establish a standardized, industry-wide approach for age verification at the device level in order to streamline app installation and usage by youth, rather than leaving verification fragmented across individual sites and services. This leads to multiple points of failure given that multiple social media companies have to store and manage this protected personal data across their servers. If a vulnerability is exploited, only a single user's data would then be compromised, instead of multiple centralized databases that contain the personal information of thousands or millions of users.

Regulators must also provide a clear definition of "age-appropriate content" as it relates to different categories of ages, and can gain insights from the historical approaches of the video game (e.g. Entertainment Software Rating Board (ESRB)) industry and film industry (e.g. British Board of Film Classification's (BBFC) and the Classification and Ratings Administration (CARA), an independent division of the Motion

Picture Association (MPA)) of America. These organizations have established detailed frameworks for evaluating content elements such as violence, sexual themes, language, drug use, and other sensitive material across different age categories. In addition, for areas that are crucial to youth safety, including content moderation and platform design, regulators should mandate that companies follow industry standards to drive consistency and effectiveness, such as those on age appropriate design offered by the Institute of Electric and Electronic Engineers (IEEE) and the CEN-CENELEC (CEN - European Committee for Standardization / Comité Européen de Normalisation; CENELEC - European Committee for Electrotechnical Standardization / Comité Européen de Normalisation Électrotechnique Workshop).

It is also essential that regulators provide incentives for positive change rather than solely utilize a punitive, fines-based approach to compliance. This can be modelled after the National Highway Traffic Safety Administration's New Car Assessment Program (NCAP) and Euro NCAP in the automobile industry, where technological innovations in safety now provide a competitive advantage and can be a brand differentiator. Relatedly, the safety standards in place in the European toy market can inform an analogous model for digital products and services as it relates to formal comprehensive testing, inspection, and certification. Regulators must also advocate for, and help support, new legislation that can address novel instantiations of criminal behaviour fostered and facilitated by new technological advances. They must also demand improvements in operational protocols by law enforcement and related investigative authorities so that online misuse or abuse prompts a systematic and coordinated response, even across jurisdictions, instead of one that is ad hoc, fragmented, and suboptimal.

Finally, regulators must provide social media platforms with clear guidance regarding identified gaps and present concrete remediation plans that incorporate specific practices, considering those previously reviewed and any new components that emerge over time. Leaving platforms to interpret vague regulations independently risks incomplete, inconsistent, or ineffective implementation of safety measures. Such ambiguity could lead platforms to either adopt a minimalist approach to compliance or implement overly broad content moderation policies that potentially infringe on fundamental rights, including children's rights as protected under the UNCRC. Models to emulate can be found with the Federal Aviation Administration's power to mandate safety fixes from aircraft manufacturers before planes can return to service, and Ofcom enforcing strict broadcasting standards and telecommunications regulations in the UK.

Implementing collaborative measures between regulators and platforms is essential to safeguard youth well-being across the current and future digital landscape. Through the SAFEST framework and the specific practices detailed below, regulators can encourage and facilitate a more secure social media environment that supports healthy youth development while minimizing the potential for various risks and harms. This approach is critically important as we move forward so that young people can benefit from all of the positives that social media has to offer, while at the same time being protected from potential drawbacks they may face.

Methodology

This research employed a comprehensive desk research methodology to analyse the current landscape of youth online safety across multiple domains. The investigation began with an extensive review of legislative frameworks, examining EU digital policy initiatives such as the DSA, OSA, AI Act, Digital Services Act, and Better Internet for Kids+ Strategy, alongside US developments including the Kids Online Safety and Privacy Act and related regulatory efforts. The academic literature review encompassed peer-reviewed research on youth development, mental health, and the documented impacts of social media on adolescent well-being. Particular attention was given to studies examining online risk exposure and cyber victimization patterns. The research also evaluated social media industry best practices related to content moderation strategies, platform-specific safety measures, the implementation of age-appropriate design principles, AI-driven decision making, and more.

Primary data sources included peer-reviewed academic papers, law review articles, academic conference proceedings, regulatory documents, platform transparency reports, industry white papers, and policy briefs from advocacy organizations. These were supplemented by secondary sources, including news articles, industry blogs, and commentary on youth online safety developments by subject matter experts. This methodological approach assisted our understanding of the complex interplay between legislation, platform practices, and youth online experiences, while identifying gaps in current approaches and measures, as well as opportunities for improvement.

In order to inform and supplement our research and analyses of youth development, online safety issues, EU legislation, and platform initiatives, we recognized the critical importance of hearing directly from young people about their lived experiences, perspectives, and concerns in social media environments. To gather these insights, we partnered with ThinkYoung, a leading EU not-for-profit organization dedicated to amplifying youth voices in policy decisions through high-quality research focused on Generations Y, Z, and Alpha.

To gather these insights, we partnered with ThinkYoung, a Brussels headquartered think-tank, research institute and non-for-profit organization focusing on young people. ThinkYoung conducts studies, surveys, focus groups and data analysis on Gen Y, Gen Z and Gen Alpha. Founded in 2009, ThinkYoung has expanded to Geneva, Nairobi, and Hong Kong, studying youth behaviors and opinions in Europe, Africa and Asia - and providing decision makers with high-quality research on key issues affecting young people. Three focus groups were designed to ensure the meaningful engagement of teenagers and young people to the on-going analyses and recommendations relating to social media use and digital technology. ThinkYoung organised and facilitated the sessions to ensure the lived experiences of teenagers are amplified as key stakeholders to the discussion. A full Focus Group report and discourse analyses, authored by Charles Howard, Head of Research at ThinkYoung, Tarquinia Palmieri, Research Officer and Giulia Cerutti, Project Officer, is available at the dedicated ThinkYoung page.

Introduction

To guide the analysis and framework development, the report is structured in four sections covering the following key research questions.

Section 1: Social Media in the Lives of Youth

1. How does social media benefit young populations?
2. What are some clear potential risks and harms associated with youth and social media?
3. What are the primary factors that contribute to excessive or otherwise unhealthy use of social media?

Section 2: Foundation of Regulation Impacting Youth In Europe

4. What is the existing state of legislation in the European Union (EU) centred on youth online safety?
5. What are the similar legislative efforts in the United States (US)?

Section 3: Current Deficiencies in Youth Online Safety Regulation

6. Where are there inefficiencies, inconsistencies, deficiencies, and overlaps in current EU legislation? How can existing legislation be synthesized, streamlined, optimized, and developed?

Section 4: A New Framework for Improving Youth Online Safety

7. Through what framework should improvements be made to addressing youth online safety in Europe?
8. What specific practices must platforms prioritize to safeguard youth online?
9. What specific practices must regulators perform to assist platforms with their compliance efforts?

To address these questions, we first review what the research has found on how certain affordances of social media meet many important needs that youth have. We then summarize the range of risks and harms that some youth may face while using various apps. Obtaining a thorough understanding of teens' online experiences is critical for informing the development of effective and targeted legislative solutions. We also look at how current views on the issues surrounding youth and social media have developed, and we analyse the assumptions behind the heated discussions on these topics.

Next, we examine existing online safety regulations for youth in Europe and the United States. This comparative analysis is essential given the extensive work being done in the EU and the heightened attention in the US surrounding the potential link between technology use and the mental, psychological, and physiological well-being of youth. This review comprehensively explains the diverse approaches to youth online safety in multiple jurisdictions. Through this analysis, we attempt to highlight potential deficiencies, issues, and areas of fragmentation among existing legislation. We then present a new model intended to arrive at the SAFEST (Safety, Autonomy and Choice, Free Expression, Evidence-based Practices, Security and Privacy, and Transparency) approach to youth online safety.

This new, streamlined, modular framework introduces specific steps for collaborative implementation by both platforms and legislators, and harmonizes their efforts as they co-labour towards the end-goal of safeguarding and supporting youth. It was built based on a consideration of the multifaceted nature of online engagement, its ubiquity and intersection with almost every area of life, the urgent need to foster positive, prosocial experiences among youth, the current mental health struggles that young people face, the concerns that governmental authorities have about the design and operation of social media companies, and the constraints within which those platforms must operate.

Section 1: The Relevance of Social Media in the Lives of Youth

Potential Positives of Social Media Use

Since the term "social media" was first used in 1994 by developer Darrell Berry who was building Matisse, a Tokyo online media environment,[1] a number of definitions have emerged. Given its evolution over the course of the last 30 years, it is helpful to use a contemporary formal definition that is high-level enough to reflect and include those changes. As such, we prefer a conception of social media as comprising "various user-driven platforms that facilitate diffusion of compelling content, dialogue creation, and communication to a broader audience. It is essentially a digital space created by the people and for the people, and it provides an environment that is conducive for interactions and networking to occur at different levels (for instance, personal, professional, business, marketing, political, and societal)".[2] In practice, it is used as an umbrella term to refer to a number of online platforms which include, but are not limited to, "blogs, business networks, collaborative projects, enterprise social networks (SN), forums, microblogs, photo sharing, products review, social bookmarking, social gaming, SN, video sharing, and virtual worlds".[3]

Social media has become deeply embedded into young people's lives, serving as a vital channel for social connection, information gathering, and identity formation. These platforms fulfil vital needs for peer interaction and relationship maintenance during adolescence, a period when social connections are vital for development.[4-6] For example, platforms like Instagram and Snapchat allow teens to share daily experiences, maintain friendships across distances, and feel connected to their peer groups even when physically apart.[4]

Of course, the role of social media in adolescent life extends beyond mere social interaction. These platforms have become primary sources for finding, accessing, and engaging with information and educational content relevant to young people's lives.[7] For instance, many teens use YouTube to supplement their formal education by watching tutorials on academic subjects or learning new skills. TikTok has emerged as a platform where young people consume entertaining content and share and gain knowledge on topics ranging from current events to mental health awareness.

Moreover, social media is a vital resource and medium for identity exploration and self-expression during adolescence. It provides

> "I think you can learn a lot of stuff by using social media, for example, there's a lot of informative videos. It also helps you stay connected with your friends."
>
> *— Female, 15 years old, Italy. ThinkYoung Focus Group 3, 15th of November, 2024.*

spaces for teens to experiment with different aspects of their identity, receive feedback from peers, and develop their sense of self.[8-11] This can be particularly important for marginalized youth who may find supportive communities online that they lack in their immediate physical environments. For example, LGBTQ+ teens in conservative areas may connect with others like themselves online, fostering a sense of belonging and self-acceptance.[12]

Similarly, racial and ethnic minority youth might find cultural affirmation and mentorship through social media groups.[13] Youth with disabilities or chronic illnesses can also build peer support networks and shared experiences online.[14] As yet another example, transgender youth can obtain vital advice, support, and community during their gender transition, assets that may not readily be available in their immediate physical environment.[15] These digital spaces allow teens from underrepresented groups to explore and express aspects of their identities that may be suppressed or unsupported in their offline worlds, potentially contributing to positive identity development and psychological well-being.[16-18]

Finally, it must be noted that social media platforms serve as powerful avenues for youth civic engagement and social activism. These environments enable young people to learn about, discuss, and advocate for issues they care about, from social justice and climate change to mental health awareness and education reform conversations.[19-22] Through social media, youth can easily access information on current events, connect with like-minded individuals, and participate in online campaigns and movements. This digital activism often translates into real-world action, with young people organizing protests, fundraisers, and community initiatives. For example, movements like Black Lives Matter and climate strikes have gained significant momentum through youth-led social media campaigns.[23-25] Furthermore, these platforms allow young voices to be heard on a global scale, giving them the power to influence public opinion and policy decisions. Thus, social media facilitates youth civic participation and empowers them to be agents of change in their communities and beyond.

> **“I think you can also get a lot of new insights from different people and cultures… you can connect with people.”**
>
> *— Female, 15 years old, Belgium. ThinkYoung Focus Group 3, 15th of November, 2024.*

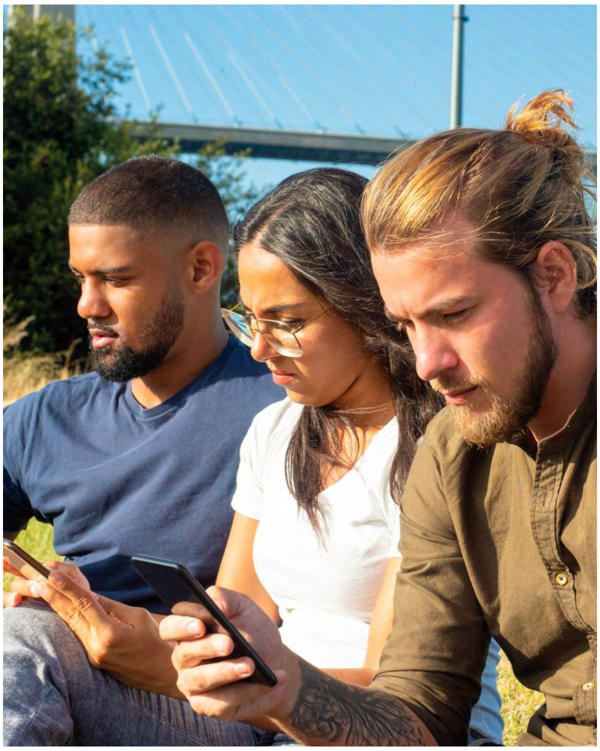
Potential Negatives of Social Media Use

Accompanying the positives of social media participation are a host of risks and harms that youth may face online, many of which have received a significant amount of attention from families, communities, educators, mental health professionals, law enforcement, paediatricians and other health care workers, and governmental officials. To systematically understand these challenges, the CO:RE (Children Online: Research and Evidence) framework offers a comprehensive classification system that organizes online risks into four distinct dimensions: Content, Contact, Conduct, and Contract risks.[26, 27] The table in Appendix A illustrates specific examples of behaviours and risks within each dimension.

While these online risks are serious concerns, it is important to note that most youth do not experience severe harm from social media use. The prevalence of these experiences remains relatively low compared to the total youth user base. That said, the evolving landscape demands stronger protective measures at both governmental and platform levels to mitigate potential emotional, psychological, and behavioural impacts. This becomes even more pressing with the emergence of extended reality technologies and generative artificial intelligence (AI), which introduce novel risks such as privacy violations in metaverse environments and AI-generated harmful content. These technological advances will continue to create new challenges in the short- and long-term and should compel proactive rather than reactive solutions to safeguarding youth online.

Factors That May Contribute to Unhealthy Use of Social Media

Beyond the aforementioned interpersonal challenges that youth may face on social media platforms, there are significant concerns about patterns of usage that can lead to psychological and emotional harm. These patterns manifest in two primary ways: overuse that interferes with daily functioning, and exposure to potentially harmful content loops. When youth become caught up in these cycles, they may repeatedly consume content that reinforces negative emotions or limited perspectives, creating echo chambers that prevent exposure to diverse viewpoints or healthier content that can truly benefit and support them. This targeted content consumption, driven by a number of personal and technological factors, is important to understand given its implications for overall health and well-being among young people.



> **“I think it depends on how much you use it. If you use it only to talk with friends, then it’s OK, but if you use it for 3 hours a day, for example, then the negatives outweigh the positives.”**
>
> *— Male, 15 years old, Belgium. ThinkYoung Focus Group 3, 15th of November, 2024.*

Ubiquity of Mobile Devices and Design of Apps

The ubiquity of mobile devices and the design of apps have significantly contributed to the prevalence of social media usage, particularly among young people. Indeed, some scholars argue that the excessive use of social media can be attributed to the omnipresence of smartphones and their role as the primary gateway to social and messaging applications.[28] The convenience offered by these devices is unparalleled, allowing users to access apps instantly with minimal effort, which has made smartphones the preferred choice for social media engagement despite the availability of other devices such as tablets and computers. The design of social media apps has been optimized to capitalize on this mobile-first approach, incorporating features that leverage the unique capabilities of smartphones. These include push notifications for new followers, comments, and private messages, location-based services, gamification elements, the ability to easily capture and share photo and video content via their camera functionality, and user-friendly filters and augmented reality experiences all helping to drive youth engagement and encouraging further usage and interaction. The seamless integration of these features into the mobile experience has further reinforced the smartphone's position as the primary tool for social media access and interaction.

Psychological Triggers

FOMO (Fear of Missing Out) also plays a role, particularly among youth who need to be "in the know" about what is happening in the news, their community, or the lives of those they follow. This phenomenon may also be relevant when considering ephemeral content that disappears after a specified time (e.g., 24 hours), as a user may feel compelled to ensure they do not miss out on information that may be valuable socially, relationally, or professionally. Variable rewards – in terms of the number of likes or comments a user receives with each new post – also may induce more frequent participation on social media than usual because of a desire to be seen, noticed, and affirmed by others.[29-31] Relatedly, it is natural to want to know if content one posts is viewed, enjoyed, and shared, and if messages one sends are received and read – contributing to increased usage of these platforms. A final component is push notifications, where incoming sounds and banners on one's phone alert them that new content they likely would be interested in is now available for them to see within specific apps. Research has shown that these external cues can increase phone and social media usage and increase overuse.[32, 33]

Neurological Triggers

Third, social media, like other activities such as gaming, shopping, and gambling[,34, 35] triggers dopamine release in regions of the brain associated with pleasure and motivation.[36] While this effect is similar to substance use, it is typically milder and does not cause severe physical health issues linked to drug addiction.[37] Social media addiction is primarily a behavioural problem, unlike drug addiction, which affects both behaviour and body functions. The widespread use and acceptance of social media can make it challenging to recognize problematic usage. However, this ubiquity might make it easier to moderate social media use compared to drug use cessation. Treating social media addiction usually involves changing behaviours and developing healthier digital habits, rather than the intensive medical treatments often needed for drug addiction.[37]

AI-driven algorithmic recommendations for content feeds significantly impact user engagement and content consumption patterns. While these algorithms can potentially lead users towards harmful content, they also possess the capability to guide users towards more positive and beneficial content. For instance, when a user is caught in a cycle of engaging with negative content like unhealthy diet tips, AI algorithms can be programmed to recognize this pattern and intentionally introduce more positive content, such as posts promoting healthy body image. This adaptive approach leverages the same machine learning techniques used to increase engagement, but to improve user well-being. Studies have shown that content-based and collaborative filtering methods can be fine-tuned to prioritize content that aligns with specific health and wellness objectives. By analysing user behaviour, preferences, and interaction patterns, AI algorithms can identify opportunities to subtly shift content recommendations towards more constructive themes, potentially breaking negative feedback loops and fostering a more balanced content diet for users.

Engagement-Centric Design Features

Engagement-centric design features in social media platforms serve a dual purpose of enhancing user experience while promoting prolonged platform usage and "stickiness." Some elements, such as streamlined interfaces, personalized content recommendations, and one-click sharing options, may provide genuine user benefits through improved relevance and usability.[38-40]

Key mechanisms like "autoplay" and "infinite scroll" eliminate natural stopping points, while strategically timed notifications prompt users to return to the platform. These features represent the natural evolution of social media companies' products and services, driven by economic incentives to maintain and expand their user base.[32, 41] Users derive benefits from reduced friction in content discovery, customizable playback options, and constant access to diverse content formats that serve various needs.[42]

The design architecture deliberately incorporates psychological principles to shape user behaviour. This includes social proof through metrics like likes, shares, and follower counts, as well as reciprocity mechanisms that encourage ongoing interactions where users feel compelled to return to the platform and engage further. While these features enhance platform functionality, they have also sparked concerns among stakeholders focused on youth mental health and well-being. The implementation of these elements reflects a complex balance between the benefits that platforms want to provide to users to ensure an enjoyable experience, and issues with engagement optimization among a vulnerable youth population in modern social media design.

It is important to note that platform design can be employed in ways that enhance user well-being. Some applications, like health, self-care, and education apps, utilize similar techniques to promote positive behavioural choices and habits.[43-45] The key difference lies in the intent and outcome of the design choices. Furthermore, it should be made clear that excessive social media use, particularly among youth, is not solely attributable to design elements. Other factors play significant roles, including the ubiquity of technology in daily life, fear of missing out (FOMO), and the brain's reward system activated by stimulating or otherwise pleasurable experiences. Indeed, these elements are relevant to other behaviours that may become excessive, including gambling, gaming, shopping, and general Internet use.[46, 47]

Evidence around the Risks of Social Media Use and Mental Health

The narrative surrounding youth mental health has increasingly centred on social media as the primary culprit for declining well-being among young people. However, this perspective oversimplifies a complex phenomenon shaped by numerous interconnected factors. While concerns about social media's impact warrant attention, focusing solely on social media platforms risks overlooking other significant variables that have profoundly affected youth mental health, including global events, socioeconomic pressures, environmental concerns, and systemic barriers to mental healthcare. A more nuanced examination reveals that the challenges facing today's youth stem from a multifaceted array of influences, each deserving careful consideration when determining the direction of regulatory efforts that will have profound impacts.

Recent Impacts on Youth Well-Being

First, while some countries like the US,[48] England,[49] and Sweden,[50] have reported an increase in mental health problems among youth across recent years, other countries like Canada,[51] the Netherlands,[52] and Norway[53] have identified stable trends.[54, 55] Across the world, there is no universal indication that the ramp-up in social media use by youth has occurred concurrently with a global decrease in youth well-being. In addition, other important variables may be at play. Research has shown mixed and inconclusive findings related to the effect of Internet use by young people and psychological health and well-being.[55-59] Furthermore, while social media use typically increases after life satisfaction decreases, it is not clear that social media use causes decreased life satisfaction.[60]

It is also inarguable that the COVID-19 pandemic had a profound and widespread impact on youth well-being globally, affecting various aspects of their lives including freedom of movement, social interactions, educational routines, economic stability, and both emotional and physical health.[61] Numerous studies across different countries have documented the pandemic's detrimental effects on young people's mental health. These impacts include increased rates of depression, anxiety, loneliness, suicidal ideation, and overall decreased life satisfaction and expectations. Such adverse outcomes have been observed in diverse nations, including Norway,[62] Germany,[63] the United States,[64] Australia,[65] Indonesia,[66] England,[67, 68] and Canada.[69] The severity of this impact is further highlighted by meta-analytic studies comparing pre-pandemic data from 2015,[70] with data collected during the pandemic in 2021[71] which indicate a substantial increase in both depression and anxiety rates among children and adolescents.[61]

Youth well-being in recent years may be compromised by various factors beyond social media use. Financial stresses experienced by families due to volatile macro-economic conditions[72] have been shown to impact children's mental health.[73-75] Additionally, increased parent-child conflict,[76, 77] and heightened caregiving responsibilities[78] contribute to the strain on young people's emotional well-being. Scholars have also posited that adolescent mental health problems are exacerbated by the broader context of growing up in an age of uncertainty[79] particularly with regard to their physical and social environments.[80] Related to this, heightened political polarization, economic instability affecting household dynamics, global health crises such as the COVID-19 pandemic, the looming threat of climate change,[81] geopolitical conflicts, school shootings, and the devastating impact of the opioid epidemic all may be contributing factors. Additionally, elements such as sleep deprivation, decreased face-to-face social interaction, intensified academic pressures, evolving family dynamics, and exposure to bullying and cyberbullying may also play a role in the observed decline in youth mental health.

As a final major point, a significant proportion of young individuals grapple with clinically diagnosed mental health disorders and may struggle to obtain professional support due to finances, stigma, or other restrictions. Moreover, many young people may struggle with subclinical symptoms or subthreshold disorders—mental health challenges that fall short of meeting the full diagnostic criteria for a clinically recognized condition. Even if they desire professional support, these youth may lack access to qualified mental health professionals or robust support networks, leaving them ill-equipped to navigate the complexities of their emotional and psychological experiences. This gap in resources and support potentially increases their vulnerability level and may impede their ability to develop effective coping mechanisms during this critical developmental stage of their life.

Collectively, these observations about the state of youth provide a nuanced perspective of their situation and highlight the intricate nature of causes and correlates. While the increase in social media and mobile device usage may loosely track with certain trends related to youth well-being, numerous other concurrent developments and flashpoint events do as well and have built a complex array of stressors that deeply affect young people today.

Research on Parental Controls

Parental controls have emerged as a seemingly promising solution for mediating children's digital engagement, appealing to both governments and the market. Parents generally wish to play an active role in their children's digital lives, despite sometimes being influenced by media-driven anxieties and a culture of parent-blaming.[82-84] However, the reality of parental control usage and effectiveness is more complex. While parental controls may seem ideal, their actual implementation and benefits are not as straightforward or universally effective as initially perceived. While many parents believe these tools can enhance their control and their child's safety online, a significant proportion are uncertain about how or whether they work.[85] Despite the availability of these tools, their adoption remains relatively low across Europe, with usage rates varying from 11% in Lithuania to around one-third in countries like NO, PL, and ES.[85] Furthermore, most parental controls fail to address the full spectrum of online risks, including content, contact, conduct, and contract risks.[86]

Recently, a comprehensive review of 40 empirical studies that have examined parental control use to facilitate children's safety online was conducted.[87] Mixed results were found across studies, with 17 reporting beneficial outcomes, 12 showing no change, 6 indicating limiting outcomes, and 8 suggesting adverse outcomes. While some studies demonstrated positive effects such as reduced exposure to online risks such as cyberbullying, sexual content, and privacy violations, some studies indicated that these tools could limit children's online opportunities, including reduced overall internet use, restricted access to beneficial online activities, and reduced privacy and autonomy for children online, and even reduced digital competence. They also may limit children's opportunities to learn about online risks, develop coping skills, and negotiate their specific needs with parents.[88]

It is important to note that most of these studies were cross-sectional, with generally small effect sizes. Moreover, the ease with which children can bypass parental controls was highlighted in some studies reporting null findings. Furthermore, both children and parents expressed concerns about increased family conflict, eroded trust, and invasion of privacy resulting from using these tools. These outcomes suggest that while parental controls may offer some benefits, they are not a standalone solution and can potentially have adverse effects on family relationships and children's digital skills development.

Indeed, research indicates that parental controls are most effective when integrated into a broader approach to parental mediation and family communication about digital engagement.[89, 90] From a child development perspective, overly restrictive or privacy-invasive parental controls can be counterproductive. Instead, these tools should aim to promote children's agency, development, safety, and privacy while preserving online opportunities. Building trust within the family is essential, as measures perceived as too restrictive or invasive can erode trust and lead to increased family conflict. The research emphasizes that warm parent-child relationships, open communication, and collaborative decision-making are more effective in managing online risks than technical restrictions alone.[91] Therefore, parental controls should be integrated into daily family practices as part of a healthy parent-child relationship, rather than relied upon as a standalone solution.

Other studies have highlighted the significant benefits of co-designing parental control tools with both parents and children. For example, Meta's Trust, Transparency and Control Labs (TTC Labs), in partnership with the design and innovation agency Smart Design, conducted extensive co-design sessions with teens and parents/guardians across multiple countries including the US, UK, Ireland, Brazil, Australia, and Japan. This initiative involved over 103 teens and 92 parents/guardians and aimed to uncover product-relevant insights and principles for developing parental supervision tools. The co-design process employed innovative methods like the "Would You Rather" game and activities such as "You Make the Rules" for teens and "Build Your Own Control Center" for parents. These sessions revealed the tensions between youth and parent priorities, informing the development of more balanced and effective parental supervision features that consider both teens' need for privacy and parents' desire for oversight.[92] Such collaborative approaches allow for the creation of more effective and balanced tools that consider the diverse needs, perspectives, and concerns of both teens and parents. We encourage more of these partnerships to foster and facilitate solutions that promote youth autonomy and address parental safety concerns without undermining the trust and relational harmony necessary for healthy families and prosocial behavioural choices.

> **"There can be an issue later because uneducated parents can end up staying away from social media and prevent their children from joining as well. Knowing how children are, that usually results in rebellious kids who install apps behing their parents backs. At the end of the day, that's not really safe because if anything happens, they're scared to tell their parents since their parents don't even know they installed that app."**
>
> *— Female, 16 years old, Cyprus. ThinkYoung Focus Group 2, 12th of November, 2024.*

Section 2: Foundation of Regulation Impacting Youth

Foundational to any youth online safety framework must be an unwavering commitment to prioritize the best interests of children when it comes to their experiences using online technologies. This commitment acknowledges that children have immutable rights in digital spaces, which must be protected and upheld by all stakeholders, including tech companies, policymakers, educators, and parents. Central to these rights is children's ability to access and meaningfully participate in the digital world, ensuring equitable access to technology and the internet and engagement with age-appropriate content and services that enhance their development, education, and social connections. Equally important is safeguarding children's user data and privacy, implementing robust protection measures, and ensuring transparency in data usage. Children also have the right to be protected from various online harms, which requires robust safety measures and moderation practices.

Moreover, the digital environment should provide educational affordances and skill-building opportunities, fostering digital literacy, critical thinking, and problem-solving abilities. Additionally, children's right to play, recreation, and leisure online must be recognized, acknowledging the importance of digital spaces for fun, creativity, and social interaction. Children should also have a voice in developing policies and technologies that affect them, participating in the design of safety features and content moderation policies. Protection from commercial exploitation, including manipulative advertising and exploitative business practices, is essential. Lastly, children have the right to comprehensive digital literacy education, empowering them to navigate the online world safely and make informed decisions. Considering and appreciating these immutable rights serves as the foundational cornerstone upon which all trust and safety efforts devoted to youth online can and must be built. These rights are not merely aspirational; they are essential, critical, and must be respected first and foremost before any policies, programs, products, or services are developed. By placing these rights at the forefront of approaches to youth online safety, companies and governments can ensure that their efforts are grounded in a fundamental respect for children's dignity, autonomy, and potential in the digital age.

The multitude of harms that can affect youth online, and the acceleration of these and new risks through AI capabilities and other emerging technologies, have prompted increased governmental scrutiny and regulation of social media platforms given concerns about their ability to regulate themselves. Requirements involving a standard duty of care, enhanced safety measures, restricting or eliminating access to platforms based on age, and requiring risk assessment and annual compliance with regulation have all been part of the public and legislative discourse in recent months and years. While these discussions and proposed measures have been prominent in many countries, the European Union (EU) has taken a particularly proactive stance in addressing youth online safety through such efforts. The EU's approach serves as a benchmark for many other regions, representing an impassioned effort to create a safer digital environment for young users. We now turn our attention to the key EU laws and regulations that have been implemented or proposed to protect children in the online sphere. The sections below detail the basis for regulation across Europe (with legislation from the US also examined for comparative purposes) in an attempt to uncover gaps, deficiencies, fragmentation, and potential unintended side effects that might emerge.

International Guidance to Consider in Regulation Impacting Children

The UN Convention on the Rights of the Child (UNCRC) has formed the backbone of thinking globally on children's rights since being adopted in 1989. However, the rapid evolution of the digital landscape has prompted a re-evaluation of how these rights apply in the online world. In response to this changing context, the General Comment No. 25 on children's rights in relation to the digital environment was published in 2021, and provided recommendations on implementing the principles of the Convention within the digital environment. It should be highlighted that the UN Convention on the Rights of the Child (UNCRC) and General Comment No. 25 both emphasize "the best interests of the child" as the primary consideration in all actions affecting children, while also recognizing the importance of respecting children's "evolving capacities." These documents highlight the key role of parents in child development and upbringing, particularly in navigating the digital landscape. They strongly advocate for children's right to freedom of expression, which extends to seeking, receiving, and sharing information and ideas through various media, including digital platforms.

In addition, the UNCRC and General Comment No. 25 both assert that children's views should be given due weight according to their age and maturity, with a recognition of their evolving capacities. They also stress the importance of implementing appropriate guidelines to shield children from potentially harmful information or material. This delicate balance acknowledges that children possess multiple (and sometimes competing) rights that must be carefully considered, balanced, and respected. An overarching principle, as stated in the Convention, is to "ensure to the maximum extent possible the survival and development of the child." This should inform all actions and policies related to children's rights, both online and offline.

How US Regulators are Approaching Social Media Use by Youth

The most significant development in the US involving governmental mechanisms to protect youth online is the Kids Online Safety and Privacy Act (KOSPA). It is a comprehensive bill that combines the Kids Online Safety Act (KOSA) and the Children and Teens' Online Privacy Protection Act (COPPA 2.0). It passed in the US Senate in July 2024, but did not make it through the US House Committee before Congress closed session for the year. A cornerstone of this Act is establishing a "duty of care." This legally obligates platforms to implement reasonable design features to prevent and mitigate various harms children might encounter while using their products and services.

Apart from the duty of care, the law requires platforms to provide minors with the highest default safety and privacy settings, create a point of contact for schools to report potential harms to minors, and provide more robust parental controls. Platforms must also limit certain design features that reward youth for staying online, and prevent age-restricted products and services from being marketed to youth. Other provisions include restrictions on sharing the geolocation data of minors, preventing unknown adults from contacting youth, and allowing minors to opt out of personalized recommendations. Moreover, platforms with over ten million monthly active users in the US must report annually on foreseeable risks of harm that minors could face, and provide updates on what steps have been taken to prevent and reduce risks.

Aside from this federal legislative effort, it is notable that some states are passing laws intended to curb the effects of addictive design elements in social media apps. In New York, the Stop Addictive Feeds Exploitation (SAFE) for Kids Act prevents platforms from providing suggested posts to teens under 18, and also prohibits platforms from sending notifications to the phones of minors between 12 am and 6 am unless parental consent is given. The Maryland Age-Appropriate Design Code Act bans platforms from using design features that encourage excessive use. In California, the Protecting Our Kids from Social Media Addiction Act (SB 976) (set to take effect in 2027) prohibits social media platforms from providing "addictive feeds" to minors without parental consent. The law also restricts notifications during late-night hours and school time, and requires platforms to offer additional parental control over their children's social media usage.

European Regulation Impacting Youth

In Europe, the EU Strategy on the Rights of the Child (RoC) was adopted by the Commission in 2021 to ensure the protection of rights of all children and secure access to basic services for vulnerable children This underpins the ethos behind various, subsequent regulations which impact children. The new strategy for a Better Internet for Kids (BIK+),[93] adopted in 2022, was designed to ensure that children are protected, respected and empowered online, in line with the European Digital Principles. It focuses on developing safe digital experiences to protect children from harmful and illegal online content, conduct, contact risks and improve their well-being online through a safe, age-appropriate digital environment created to respect children's best interests.

Throughout the UK and the EU, regulation has emerged that has targeted one of the following areas:

1. **Addressing Illegal and Harmful Content:** This includes restricting access and/exposure to harmful, age-inappropriate, or illegal content through processes such as conducting risk assessments and implementing content moderation practices;

2. **Transparency and Due Process:** This includes mandating transparency reports to increase accountability and understanding of platform practices. Some regulation also mandates new types of due process to ensure fair outcomes for users;

3. **Privacy and Security:** This includes protecting personal data of minors and avoiding profiling of minors for advertising purposes, mandating highest privacy settings by default for children, and protecting children's security (from things like phishing, scams, identity theft, and other security threats);

4. **Child Sexual Abuse and Exploitation:** This includes protection from all forms of sexual violence, such as child sexual abuse and exploitation;

5. **Use of AI and Automation:** This includes preventing addiction and negative mental health impacts through AI "rabbit holes" and thoughtfully considering algorithmic recommendations and use of AI to support content moderation;

6. **Age Assurance:** Enforce policies based on knowing the age of a user on the platform through "highly effective" methods.[94]

Addressing Illegal and Harmful Content

The Digital Services Act (DSA), enacted on August 25, 2023, implemented new rules and obligations to safeguard EU citizens as they use online platforms. When it comes to tackling illegal content, the DSA does not aim to define or amend content legality. What constitutes illegal content is defined in other laws either at EU level or at national level (e.g. terrorist content, child sexual abuse material, or illegal hate speech). The DSA asserts that where content is illegal only in a given Member State, as a general rule it should only be removed in the territory where it is illegal.

However, the DSA aims to establish an EU-wide framework to better detect, flag and remove illegal content, as well as new risk assessment obligations for very large online platforms and search engines to identify how illegal content spreads on their service. For example, it requires platforms to have easy-to-use flagging mechanisms for illegal content. Platforms should process reports of illegal content in a timely manner, providing information to both the user who flagged the illegal content and user who published the content on their decision and any further action. A priority channel will be created for trusted flaggers – entities which have demonstrated particular expertise and competence – to report illegal content to which platforms will have to react with priority.

When it comes to harms which impact children, platforms are expected to "put in place appropriate and proportionate measures to ensure a high level of privacy, safety, and security of minors, on their service" with the best interests of the child in mind. This includes a requirement for platforms to consider the risk of minors finding content that could harm their "health, physical, mental and moral development" ("age-inappropriate content"). The 5Rights Foundation has outlined ways that platforms can assess this through tools such as Child Risk Impact Assessments (CRIAs), discussed further, alongside other actions platforms can take in Appendix B.

In the UK, the Online Safety Act (OSA), enacted on Oct 26, 2023 also has a large focus on restricting illegal material and content that is harmful to children. Companies will now need to prevent, detect and remove illegal content, which includes content depicting, promoting or facilitating child sexual abuse, terrorism and suicide amongst other priority offenses. This includes image-based sexual offences (including possession of extreme pornography and non-consensual disclosure of intimate images). Service providers in scope of the OSA are required to take down content where they have "reasonable grounds to infer" that content is illegal.

While there are similarities between the DSA and the OSA, the content in scope under the latter is more specific regarding what minors should not be able to access. Beyond illegal content, the OSA also states that the following three types of harmful content should be protected against:

1. Primary priority content that is harmful to children (pornographic content or content that encourages/promotes/instructs on suicide, deliberate self-injury or eating disorders/behaviours).

2. Priority content that is harmful to children (bullying content or content which (a) is abusive and targets race, religion, sex, sexual orientation, disability or gender reassignment (b) incites hatred against people based on these characteristics, (c) encourages/promotes/instructs on serious violence against a person or a challenge/stunt likely to result in serious injury, (d) depicts serious violence or (in graphic detail) serious injury against a person/animal/fictional creature, (e) encourages self-administration of physiclly harmful substances).

3. Content not within (1) or (2) that presents a material risk of serious harm to an appreciable number of children in the UK (non-designated content that is harmful to children).

Both the DSA and OSA mandate risk assessments, however the DSA focuses on assessing and mitigating systemic risks arising from the design and functioning of platforms. It requires Very Large Online Platforms and Search Engines (VLOPs and VLOSEs) to conduct annual risk assessments to measure any negative impact of their service on children's rights, and to assess how the features built into the design of the platform may cause addiction.

The OSA risk assessments focus on protecting against exposure to illegal content. The risk assessment required by the UK's Office of Communications (Ofcom) mandates that platforms consider factors such as the make-up of its user base, how algorithms used by their service contribute to the spread of illegal content and how easily, widely, and quickly illegal content can spread; the level of risk of your service being used for the commission of a priority offence; the level of risk of harm to individuals presented by illegal content; how the functionalities of your service (e.g. direct messaging) facilitate the presence or dissemination of illegal content; how the way individuals use their service contributes to the spread of illegal harms; the nature and severity of the harm that could be suffered by individuals due to illegal content present on the service; and how the design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified.

In addition to the mandatory illegal content risk assessment in scope for any company following under the OSA purview, Ofcom has provided guidance on conducting children's risk assessments, structuring this into a 4-stage process where entities: (1) understand the harms, (2) assess the risk of harm to children, (3) decide measures, implement, and record and (4) report, review, and update risk assessments.[95] Companies may tailor their risk assessment based on the specifics of their platforms (and the features that their platform enables); however, the guidance from Ofcom provides useful considerations to take into account, such as how features and functionalities affecting frequency of use increase risk of harm, the platform's business model, and other such considerations. It also provides further detail on risks within each of the harm areas identified.

In addition, companies must assess the risks and dangers that their platforms pose to the safety of children. If risks are identified, companies are required to act by implementing mitigation strategies. Larger companies (defined as one that has more than 7 million monthly UK users, or roughly 10% of the UK population) will also need to publish a summary of their risk assessments in an effort to promote increased transparency around the risks that online platforms and services may pose to children.

Transparency and Due Process

The DSA involved a number of requirements to enhance transparency. For example, platforms must be transparent about their terms of service; they must make sure their terms and conditions are easy for minors to understand and consider risks that might emerge if they cannot grasp how the platform works. Platforms will be required to produce annual reports on their content moderation efforts, including the number of orders (received from Member States or "trusted flaggers") to take down illegal content, as well as the volume of complaints from users and how these were handled.

The DSA also includes requirements for providers of online platforms to maintain an internal complaint handling system and to engage with newly established out-of-court dispute settlement bodies. The internal complaint system enables users to make complaints against restrictive decisions taken by an online platform as a result of a notice (such as removing content, restricting visibility of content, or other such actions). Under the DSA, out-of-court dispute settlement bodies offer an additional opportunity for users to resolve content moderation disputes with online platforms, providing an independent third-party to support due process rights for users. These bodies will not be empowered to impose binding decisions and users can still initiate proceedings before a court if they want to contest decisions made by providers. Importantly, all these provisions are subject to reporting requirements on how well they are functioning. Intermediary services will need to report on the number of orders they have received from Member state authorities; Trusted Flaggers and Out-of-Court Dispute Settlement Bodies will be subject to their own transparency reporting requirements; and European and national regulatory bodies will need to conduct regular assessments.

The OSA also significantly increases transparency requirements for platforms. Platforms must undergo independent third-party audits and publish detailed reports about their safety measures, risk assessments, and content moderation practices.

These reports must reveal specific data about illegal content prevalence, user exposure to harmful material, and the effectiveness of child protection features. Ofcom will analyse these transparency reports to identify best (and worst) practices across the industry to inform users as they consider the safety of various platforms. Failure to comply with these transparency requirements will result in substantial penalties of up to £18 million or 10% of worldwide revenue.

Privacy and Security

In Europe (and arguably globally, given its adoption), one of the most prominent privacy-centric regulations is the General Data Protection Regulation (GDPR), which, since August 25, 2018, has endeavoured to safeguard children's personal data by mandating parental consent for processing information of those under 16 (or a lower age set by member states). In addition, the ePrivacy Directive in the EU makes sure that all users (including children) can use electronic communications in a confidential way and that their devices are protected. While the GDPR focuses on personal data protection, the ePrivacy Directive aims to protect the privacy of EU residents' electronic communications content and its metadata, even where that includes non-personal data.

The DSA aims to complement the rules of the GDPR to ensure the highest level of data protection. When it comes to handling personal data in advertising, both the DSA and GDPR regulations apply to platform service providers. In addition to the GDPR requirements for any personal data processing, the DSA prohibits targeted advertisements by online platforms using user profiling that relies on the special categories of data specified in Article 9(1) of the GDPR, such as sexual orientation, ethnicity or religious beliefs. Platforms are also prohibited from targeted advertising practices towards minors (i.e., anyone under 18), and data harvesting for profiling purposes. This is even in the case when the providers are aware with reasonable certainty that the user is a minor. According to the DSA, online platforms used by children should protect the privacy and security of their users. While there is no formal guidance on how to achieve this, child-rights organizations emphasize several essential principles for protecting young users in digital spaces.[96] At the core is data minimisation, which insists that companies collect only the minimum amount of personal data necessary to fulfil a specific purpose. Furthermore, privacy protection must be built into the system's foundation, with high-privacy settings configured as the default for all young users. These settings should automatically restrict the visibility of children's accounts, limit their exposure to potentially harmful contacts and content, and include robust mechanisms for the swift removal of illegal material.

The UK's Information Commissioner's Office (ICO) released an Age Appropriate Design Code[97] (now the Children's Code) that has been in force since 2020 and highlights how General Data Protection Regulation applies to children using digital services. It specifically mandates "high" privacy settings for all youth accounts across digital services and platforms by default. Said another way, children's information is not visible or accessible to other users without explicit action from the child to enable such access. Additionally, the Code mandates that any optional features that use personal data, including personalization services and third-party data sharing, must be individually activated by the child rather than automatically enabled.

Combatting Child Sexual Abuse and Exploitation

In the EU, the Directive 2011/93/EU on combating the sexual abuse and sexual exploitation of children harmonizes the approach to tackle Child Sexual Abuse Material (CSAM) across all EU countries. This directive requires EU Member States to take a number of measures to prevent and combat child sexual abuse, including:

- Criminalising all forms of sexual exploitation and sexual abuse of children, including the possession, distribution, and production of child pornography
- Providing support and assistance to victims of child sexual abuse, including access to medical and psychological care, legal aid, and other forms of support
- Establishing reporting and referral systems to enable the identification and referral of children at risk of sexual exploitation and abuse, and
- Ensuring that perpetrators of child sexual abuse are brought to justice, including through the use of effective law enforcement measures and the provision of adequate training for law enforcement and judicial authorities

Adopted on February 6, 2024, the Recast of Directive 2011/93/EU aims to address gaps in this Directive and criminalizes all forms of child sexual abuse and exploitation, especially new forms of online child sexual abuse and exploitation enabled or facilitated by technological developments, (e.g., live streaming of child sexual abuse, deepfakes, online solicitation, sexual abuse in virtual reality settings, operation of an online service for child sexual abuse or sexual exploitation) in all Member States.

Given the ePrivacy directive, there has been a need to carve out a targeted exception for companies to scan for CSAM voluntarily. Parliament reached a provisional agreement in early 2024 to extend the interim regulation on a temporary derogation from certain provisions of the ePrivacy directive for voluntary detection of online CSAM until 3 April 2026, allowing streaming and video media applications to voluntarily detect, report and remove CSAM.

The EU is also working to ensure that companies' voluntary measures to find CSAM will be harmonized. However, the future of long-term legislation in the EU enabling platforms to tackle CSAM is currently uncertain, considering the ePrivacy directive and the contrasting views on how to be proactive in protecting children (e.g., through voluntary scanning of the platform for CSAM) while upholding important privacy goals. While the Parliament already has a position on the proposal for permanent rules to combat and prevent child sexual abuse online (one that balances detection goals while avoiding generalized monitoring of the internet activity), the Council has yet to agree on its negotiating mandate. Complicating this is the development from June 2024, where a vote on amending the draft law covering the scanning of CSAM was cancelled amidst reports that some member states were expected to abstain or oppose the law over cybersecurity and privacy concerns. There is now a lack of clarity as to the future of legislation focused on tackling CSAM.

In the UK, as Ofcom is taking a phased approach to producing and consulting on codes and guidance for companies on complying with the new duties, the DCMS and Home Office have published voluntary codes on tackling online child sexual exploitation and abuse online. The UK Home Office has also provided voluntary principles to tackle online sexual abuse and exploitation of children.

Artificial Intelligence and Automation

Automated decision making and use of AI is covered through multiple legislation in Europe. Article 22 of the GDPR and the UK GDPR gives people the right not to be subject to solely automated decisions, including profiling, which have a legal or similarly significant effect on them. This has long been the crux of Europe's legal approach to automated decision-making.

With the DSA, platforms must now clearly lay out how their content moderation and algorithmic recommender systems work in their terms of service, and they must offer users at least one option for an alternative recommender system (or "feed") not based on profiling. In the required transparency reports, platforms must also describe any automated systems used to moderate content and disclose what their accuracy and possible error rate could be. They must also give users clear information about why they were targeted with an ad and how to change ad targeting parameters.

Beyond this, the EU's Artificial Intelligence Act was voted into law by the European Parliament in March 2024; the Act explicitly prohibits using any AI system that exploits vulnerabilities related to age, disability, or socio-economic circumstances to distort behaviour, causing significant harm. The Act mostly groups children together with other 'vulnerable groups' and vaguely mandates the protection of such groups by making sure the AI system 'addresses their specific needs.' Owners of AI systems designated as "high-risk" must also take particular account of children through a detailed risk management review and build-in human oversight and data governance. Guidance on how to create (or monitor) AI tools used by children is covered more extensively by non-binding policies, including the UNICEF Guidance on AI and Children 2.0.

Age Assurance

When it comes to age assurance, there is no obligation to assess age in the DSA. Service providers can ensure they are compliant by ensuring a high level of privacy, safety and security for all users. However, if a provider chooses to offer a service that does not meet this bar and is not appropriate for children, the provider should ensure it is not accessible to them, by means of appropriate and proportionate age assurance measures.

This is in contrast with the OSA where all user-to-user and search services regulated under the Act must carry out children's access assessments. Ofcom, the UK's online regulator overseeing enforcement of the OSA, has stated that age verification or age estimation is the only way companies can conclude that it is not possible for children to access their services. UK companies are also legally required to use age verification or estimation tools to prevent children from encountering harmful or age-inappropriate content. The age assurance technology should be highly effective, technically accurate, robust, reliable, and fair. Examples of age assurance methods that have the potential to meet these criteria include photo-ID matching, facial age estimation or reusable digital identity services. Examples of age assurance methods that are not highly effective include payment methods that do not require the user to be over 18 and simply stating in a company's terms and conditions that the service is for over 18s only.

Other Regulatory Efforts

Beyond the regulation highlighted above, the European Commission has adopted a Recommendation on developing and strengthening integrated child protection systems in the best interests of the child on 23 April 2024. It encourages a multidisciplinary approach to child protection, where educators, health professionals, law enforcement, companies, authorities at different levels, and other key stakeholders work collaboratively for offline and online child protection in a systemic and complementary way. This is likely to shape the future direction, initiatives, and legislation emerging out of Europe in the years ahead. Inclusion of children in the consultation and even co-creation of relevant projects to protect children will progress, through initiatives like the new EU Children's Participation Platform, which works at EU and national levels to provide children with a safe space to have their say in important decisions.

Section 3: Current Deficiencies in Online Safety Regulation Impacting Youth

When considering the current and forthcoming landscape of platform governance through federal legislation, certain challenges, deficiencies, and gaps have emerged. Some of the regulation is based on an incomplete understanding of the risks of online engagement to youth, as discussed above, while other regulations have emerged largely because of social and political pressure (i.e., a need to demonstrate action) in the public eye, therefore, it is unclear whether this regulation will serve the purpose as envisioned. Below, we detail major concerns with these legislative initiatives before suggesting a more optimal approach that addresses those inherent limitations.

Restrictive Approaches to Digital Engagement Not Informed by Research

It is worth a deeper dive to explore a potential fallacy behind the legislation focused on social media use by youth: the assumption that if platforms simply provide more controls to parents and create more restrictive, controlled spaces in which to interact, the mental health epidemic would measurably recede. Researchers and experts have already noted the multifaceted complexities associated with the state of youth today, and so such a belief is reductionist at best. There is a need to augment the ability of parents, families, youth organizations, communities, and other institutions to meaningfully come alongside teens and give them what they need: a comprehensive support system that fosters critical skills (self-control, self-regulation, self-awareness, empathy, resilience, digital citizenship, digital literacy, emotional intelligence, and resilience). This approach should encompass education on healthy online behaviours, critical thinking skills to navigate interpersonal challenges, and the development of robust offline support networks.

In the United States, prohibitionist approaches have historically failed to work,[98-102] lack clear scientific backing,[103-106] are often circumvented,[107] and violate the right to free expression and access to information.[108] As an example, society has spent many years focused on the purported relationship between violent video games and violent behaviour, and legislation was created to safeguard children by preventing distribution of video games with certain violent content to minors. However, the courts soon granted injunctions against certain state laws in Oklahoma, Illinois, Michigan, Minnesota, California, and Louisiana largely because the research was either missing, weak, or inconclusive. Even the renowned American Psychological Association (APA) stated in 2020 that there is “insufficient evidence to support a causal link”.[109]

What is more, the APA articulated that “Violence is a complex social problem that likely stems from many factors that warrant attention from researchers, policymakers, and the public. Attributing violence to violent video gaming is not scientifically sound and draws attention away from other factors.”[110] Government reports from Australia[111] and Sweden[112] also conclude that clear empirical evidence related to this proposed relationship is lacking. Numerous cross-sectional and longitudinal studies indicate that playing violent video games does not increase the likelihood of aggression or violence.[113-116] Relatedly, research involving over one thousand British youth between the ages of 14 and 15 did not find a positive association between recent violent video game play and self-assessment of aggressive behaviour. Beliefs about an expected link may persist because of confirmation bias, selective reporting, selective attention,[117] and the precautionary principle.[113] where policymakers err on the side of caution and put protections in place when there may be a risk, even if it has not been shown empirically as of yet.

Lessons can also be learned from what we know about excessive video gaming practices. Research has identified that extreme overuse is linked to various adverse outcomes, including poorer interpersonal relationships, impaired school or work performance, and neglect of hygiene and other personal needs.[118, 119] Furthermore, since many games are immersive in nature, many gamers spend a significant number of hours per week playing but relatively few experience problematic consequences.[120, 121] This fact seems to suggest that those resultant negatives occur because of the characteristics of the individual gamers as compared to the games themselves.[106, 118]

As has been the trend, governmental regulation emerged as a potential approach to tackle problematic video gaming among youth, drawing parallels with strategies implemented to combat substance abuse and addiction in various nations.[122] South Korea’s “Cinderella Law” of 2011 is a notable example, where the national assembly, concerned about the impact of social media and gaming on children’s mental health, sleep quality, and academic performance, implemented a ban forbidding children aged 15 and younger from using the internet between 12am and 6am. Violating this law could result in penalties of up to two years imprisonment or a fine of approximately $9,000 USD. However, research showed that this decade-long ban had minimal impact, reducing internet use by less than five minutes in the first two years (returning to baseline within three years), improving sleep duration by only two minutes per night, and having no effect on test scores.[123] The ineffectiveness of these regulations in South Korea relates to various factors, including youth still being able to use non-online games and apps after hours, log in with ID cards belonging to parents or older siblings/friends, or prioritize use during the other eighteen hours of the day.

Research from various countries has also demonstrated that nationwide restrictions are ineffective in curbing excessive video gaming among youth. Similarly, China's restriction limiting under-18s to 90 minutes of daily gaming (or 3 hours on public holidays) failed to decrease "heavy gaming" (defined as playing for over four hours a day, six days a week), as revealed by an analysis of 7 billion hours of gaming over 22 weeks from 188 million gamer profiles in late 2019 and early 2020.[124] Unexpected consequences also resulted in other contexts; an experiment among gamers in South Africa identified that users were left upset and unsatisfied after being forced to stop their gaming activity at a specific time, intensifying their desire to play more immediately.[125] When considering these less-than-desirable outcomes in the past, it appears that bans and restrictions may not hold much promise. However, current interest and momentum in society remains high, as Australia passed a new law on November 28, 2024 to ban youth under the age of 16 from using major platforms such as TikTok, Instagram, Snapchat, Reddit, Facebook, and X (while excluding YouTube, WhatsApp, and various educational platforms like Google Classroom). This law is expected to take effect in late 2025, in an attempt to give social media platforms time to develop appropriate age assurance systems to comply with this law.

The DSA primarily emphasizes children's rights to safety, privacy, and security, while giving less attention to their fundamental rights of expression and participation. A holistic approach grounded in the Charter of Fundamental Rights of the European Union should balance both risk prevention and the promotion of children's assets and rights. A predominantly restrictive approach may fall short of achieving its intended outcomes; as such, implementing regulations in a way that both enhances child participation and addresses risks and harms is more likely to yield sustainable, realistic outcomes—acknowledging both safety concerns that can arise, as well as the legitimate and important desires of youth to engage and connect online.

Lack of a Systematic, Data-Informed and Evidence-Driven Approach

In addition, much of the legislation currently in place or proposed in Europe and globally lacks a substantive evidence base to support its directives. The 2024 BIK Policy Monitor Report[126] assesses the state of digital policies based on the recommended measures of the European Strategy for a Better Internet for Kids (the BIK+ strategy)[93] against the background of significant changes in the legislative and regulatory landscape across all 27 EU Member States, Iceland, and Norway. From a data collection perspective and evidence-based approach to regulation, the study found that only 8 out of the 29 countries report regular data collection on children's digital activity. About half (14 countries) indicated that there is limited or no data collection at the national level on what youth do online, and only 8 out of 29 countries report systematic monitoring and evaluation of their policies on this topic (Cyprus, Germany, Hungary, Italy, Luxembourg, the Netherlands, Portugal, and Romania). Indeed, most countries state that policies are monitored and evaluated ad hoc but not systematically. This points to a gap in the current regulatory environment in evidence-based policy formation based on robust, recent, and representative data sets. Up-to-date empirical evidence about children's digital engagement must inform policies to make them more effective.

Lack of Standardization of Safety Practices Across Platforms

The digital safety landscape for youth varies considerably across social media platforms, with significant disparities in user controls, parental supervision features, educational resources, formal policies, and content moderation practices. This fragmentation stems from companies historically approaching trust and safety issues independently, with minimal inter-platform coordination and varying resource allocations. While acknowledging these differences, legislative initiatives should establish a universal baseline of essential protections and verification mechanisms across all platforms. Doing so would eliminate inconsistencies in youth online experiences and create more predictable and safe online experiences. By establishing universally agreed-upon baseline protections, platforms can collectively contribute to youth well-being while maintaining their unique features and competitive advantages.

As just one example, currently, there are not universally accepted content moderation standards established by regulators that are mandated to apply equally across social media platforms. Whilst the DSA and OSA address parts of the content moderation practices undertaken, they do not comprehensively cover all relevant aspects that should be regulated. For example, standards that guide acceptable accuracy levels, training, moderator well-being, use of outsourcing, or other important aspects of content moderation decisions, would enable adherence to a minimum safety baseline when it comes to content moderation across the industry. This could follow similar models to regulation of physical toys in Europe (discussed further in the next section). Whilst standards exist in some areas such as age-appropriate design, age verification, tackling AI-generated CSAM, and others, none of these are mandated by regulators uniformly across Europe for platforms to adhere to.

Disregard for the Reality of Access Circumvention

Legislation limiting youth access to major social media platforms presents numerous challenges. These restrictions potentially infringe upon children's digital rights, specifically Article 31 of the UN Convention on the Rights of the Child, which recognizes every child's right to play—inclusive of social media engagement. Historical precedents demonstrate the inefficacy of prohibitionist approaches, as evidenced by unsuccessful bans on substances and activities such as video gaming. Similar restrictions on social media access may yield comparable results. Youth often circumvent such bans by utilizing older siblings' accounts, employing Virtual Private Networks (VPNs) or proxies, or migrating to lesser-known platforms with potentially fewer safety measures.

Additionally, youth experiencing issues on social media may be less likely to seek help from parents or guardians if social media is seen as something dangerous to be avoided rather than a tool that can support their personal and professional growth. This approach also fails to prepare young people to engage responsibly and intelligently with social media and technology when they are older. Such legislation may also disincentivize tech companies from designing youth-friendly products, services, environments, and controls. Lastly, and as mentioned earlier, youth from marginalized or minority groups who rely on social media to find community, support, and hope may be denied this opportunity to meet the basic human need for connection and belonging. For some young people, social media and online technologies are a lifeline. However, legislators and governing bodies appear to overlook or underestimate the critical role these platforms play in supporting vulnerable youth, potentially jeopardizing a vital resource for those who need it most.

Lack of Monitoring Structure to Measure Effectiveness of Policy Interventions and Regulation

The increasing scrutiny of social media companies and digital platforms has increased demand for clear, measurable indicators of trust and safety improvements. As governments and regulatory bodies propose and implement new laws to induce meaningful trust and safety enhancements, there is a pressing need for Key Performance Indicators (KPIs) that can effectively gauge the success of these initiatives. With current regulation established in Europe, particularly with the DSA, measurement frameworks (or KPIs) to evaluate the effectiveness of the regulation have not been established and/or publicly shared. Therefore, it is unclear as to if/how the success of this regulation is intended to be gauged. However, some progress has been made when examining the suggested implementation of the Online Safety Act (OSA).

In 2024, Ofcom partnered with State of Life (SoL), a social impact think tank in the UK, to determine an optimal action plan for evaluating youth online safety measures. In its feasibility report, the SoL team suggested that Ofcom should focus on using four personal well-being questions from the Office of National Statistics (which include measures on life satisfaction, worthwhile activities in one's life, happiness, and anxiety), as well as other existing "domain-specific" questions that connect well-being to online activity and safety in user surveys.[127] While establishing causality would be challenging, it was recommended that Ofcom monitor well-being trends before, during, and after the OSA's phased rollout to assess the impact of the regulation. Finally, they propose a dual application of well-being measures as both an outcome indicator of online harms and a predictive tool for identifying at-risk youth. This approach could help pinpoint children in particularly vulnerable groups who may be more susceptible to online dangers or prone to experiencing heightened negative effects when exposed to harmful content.

By leveraging these insights, companies can refine their educational initiatives and content moderation strategies, creating more targeted and effective safeguards for diverse youth subpopulations on their platforms. Ofcom's consideration of well-being metrics aligns with broader international efforts to measure societal progress beyond traditional indicators. We strongly recommend that all countries work with third-party researchers to determine what validated measures to use to assess the complex interplay between online interactions and youth well-being and implement systems in place to collect and evaluate these data yearly.

Fragmentation and Inconsistencies within and across Europe

In the EU, most countries have either enacted or are actively drafting legislation to give effect to the DSA and other EU laws. Six countries report that they have codes of conduct at the national level to accompany such laws (Germany, Ireland, the Netherlands, Portugal, Romania, and Sweden). In addition, 16 countries have codes of practice that seek to protect children as young digital consumers. Laws against intimate image abuse and cyberbullying are reported to be widely available. To that end, 26 countries report that there are relevant laws or policies in place to address intimate image abuse, while 20 countries report laws and policies to address cyberbullying. These refer to laws and policies in place before the DSA entered into full effect on 17 February 2024.

However, the 2024 BIK Policy Monitor Report,[126] highlighted a number of issues with the implementation of the strategy thus far, including fragmentation and inconsistency across member states. For example, only six out of 29 countries stated that a central ministry office, public agency, or regulatory authority is formally mandated to lead on and develop BIK-related policies (Cyprus, Croatia, Ireland, Italy, Norway, and Portugal). Over half report that policy development occurs across a range of ministries and that there is no lead ministry or agency with a specific assigned responsibility. Eight of the 29 countries say there is a clearly defined mechanism to coordinate cross-cutting policy issues and stakeholder involvement on BIK-related issues (Cyprus, Hungary, Ireland, Iceland, Italy, Norway, Portugal, and Slovakia). One-third, or ten out of 29 countries, report a national action plan on BIK-related topics with defined timelines, assigned responsibilities, and KPIs (Cyprus, Hungary, Ireland, Iceland, Italy, Malta, Portugal, Romania, Slovenia, and Slovakia) in place.

When it comes to implementing the Digital Services Act, some member states are clearly lagging in their enforcement plans, leading to a fragmented approach across the EU. In April, the European Commission decided to open infringement procedures by sending letters of formal notice to six Member States where significant delays in the designation and/or empowerment of their Digital Services Coordinators were expected. At that time, Spain, Poland, and Slovakia still had to designate their Digital Services Coordinators (DSC). In addition, despite designating their Digital Services Coordinators, Cyprus, Czechia, and Portugal still have to empower them with the necessary powers and competencies to carry out their tasks, including the imposition of sanctions in cases of non-compliance. In the meantime, Estonia and Slovakia have formally designated and empowered their Digital Services Coordinators.

On 25 July 2024, the European Commission decided to open infringement procedures by sending letters of formal notice to six additional Member States, namely Belgium, Spain, Croatia, Luxembourg, the Netherlands, and Sweden, for similar delays. As of December 16, 2024, Belgium, Bulgaria, Spain, The Netherlands and Poland still have to designate or empower their Digital Services Coordinators with the necessary powers and competencies to carry out their tasks, including the imposition of sanctions in cases of non-compliance.

Even once all DSCs are designated, there is a large variance in the types of entities that have been assigned this role; most EU states have added the DSC portfolio to pre-existing roles such as within telecommunications regulators or consumer protection authorities. These regulators agencies may lack the expertise and understanding of the multifaceted nature of children's rights online. As highlighted by child safety experts:

> *"That inconsistency promises inconsistency and delay in regulatory practice across the EU and potentially places an undue burden on regulators who are educated in children's online protection, practices, and rights. The law's implementation would benefit from the attention of the European Board for Digital Services (composed of member states' DSCs) to the challenges this diversity of backgrounds and knowledge represents and educate their members accordingly, with youth participation included. Research and coordination are needed to ascertain what DSCs need to know and where training is needed to effectively implement the child-focused aspects of the DSA."*[128]

In order for this regulation to be effective, the European Commission should ensure more uniform implementation and appropriate expertise across member states.

Lack of Clarity on Balancing Tension between Privacy and Safety

In many areas of digital regulation, there is not clear or effective guidance on balancing privacy and safety goals when those are in tension. This has been the case in multiple areas. The first is the debate about the use of age verification technologies. This is important for a multitude of reasons. The first is because real-world data shows that many underage users are on social media. Research by Ofcom found that fifty-one percent of children aged under 13 – which is commonly the minimum age requirement for many social media platforms – report using social media sites/apps.[129] They also identified that forty percent of 8-17 year olds admit to having provided a fake age to obtain access to a new site or app. Therefore, age verification and enforcement is important given the ability to otherwise quite easily circumvent age restrictions.

Secondly, in order to protect children without infringing on the rights of adults, there needs to be an effective way to either assess, estimate or verify the age of individuals on a platform. This is a major policy issue in many countries and with new legislation being proposed/developed. In the EU, four countries (Germany, Denmark, Estonia, and Lithuania) state that a policy on age verification is in place, while eleven countries indicate that this is actively under development. Just under half of those countries surveyed (14 out of 29) have digital identity systems for minors. However, these are generally available only to those over 14.[126]

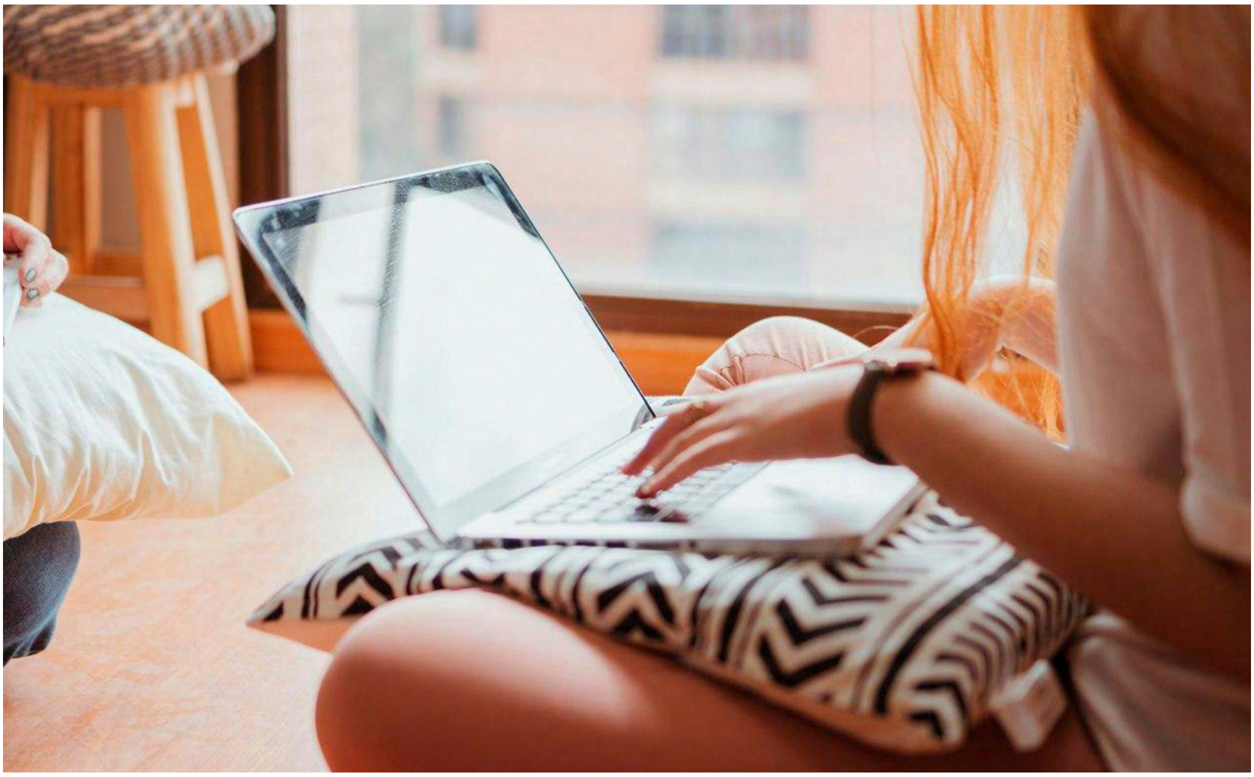
Many stakeholders believe that age verification in its current form(s) poses too large an intrusion on privacy. Others highlight various advances in technology approaches such as use of zero-knowledge proofs to maintain privacy whilst passing age checks to platforms. The Digital Trust and Safety Partnership has conducted an analysis of various age assurance methods and highlighted the challenges and best practices of different approaches.[94] Many stakeholders have argued that age verification is necessary given the significant risks of accessing apps not intended for children (or children under a certain age). There is not a single, cohesive legislative framework around age assurance across Europe and the best approach to balance multiple goals of preserving privacy whilst ensuring effectiveness. Many child-safety groups highlight the role that device-level age verification can effectively play in addressing privacy and other potential risks (discussed in Section 4 of this report).

Lack of Clarity in Legislative Language

The effectiveness of youth online safety legislation will always hinge on the measurability of its requirements. Vague language, such as protecting youth users against content "of the kind which is considered to present a material risk of significant harm to an appreciable number of children" (from the OSA), leaves too much room for interpretation and potential misapplication. One concern has to do with the effect of moral panics, which can drive policy in counterproductive ways. Over the years, various moral panics have emerged, such as the Momo Challenge and the Skibidi Toilet Syndrome[130, 131] causing alarm among youth-serving adults despite lacking substantial evidence of broad participation. Similarly, the media attention and political energy devoted to the perceived threat of child predators lurking in chatrooms and on social media sites was often disproportionate to the actual prevalence of such incidents.[132, 133] These examples underscore the importance of specificity in legislative wording, because platform actions must be centred on preventing historically proven risks and harms to youth instead of feeling obligated to respond to hoaxes, sensationalistic but short-lived trends, or anecdotal events that flare up in the zeitgeist but then quickly fade out because there was no true threat with significant, measurable impact. Said another way, platforms need to avoid reactionary policies driven by unfounded fears or exaggerated worries.

Moreover, attempting to respond to vague requirements in EU online safety laws creates an undue burden on social media platforms, as they are left to interpret ambiguous expectations. This lack of clarity can lead to several challenges worth mentioning. First, it forces platforms to make subjective judgments about content moderation, potentially resulting in inconsistent enforcement across different services. This inconsistency can confuse users and create an unevenness in how companies operate in EU countries. Second, vague requirements may lead to overzealous content removal as platforms err on the side of caution to avoid potential legal repercussions. This contravenes the fundamental goal of upholding children's digital rights by infringing upon the freedom of expression and access to information that youth must be afforded. Third, ambiguity in the language of regulations can hinder innovation and investment in the digital sector, as social media platforms may be hesitant to introduce new features or services without clear compliance guidelines. This uncertainty can stifle technological advancements that might greatly support youthful populations, and lead to confusion in where resources should be best allocated to accomplish trust and safety goals.

To address these issues, EU lawmakers should strive to provide more precise definitions and more explicit guidelines in online safety legislation. Clear, empirically measurable requirements guard against misinterpretation and overbroad application and provide platforms with definitive targets for their prevention and response actions. This enables tech companies to allocate their time and resources efficiently and effectively, rather than burdening them with ambiguous requests that may lead to ineffective measures or unintended consequences.



Section 4: A New Framework on Youth Online Safety Regulation Across Europe

The SAFEST Model

To serve as a template for Youth Online Safety regulation across the EU, we have built a model consisting of six critical components that must remain top-of-mind when considering social media platform operations and their impact on youth development, well-being, and digital rights. This, we believe, is the most optimal way to balance protection with empowerment to ensure that youth can safely participate in digital spaces without being restricted in ways that hamper their development and skill-building. This model considers both the risks and opportunities of online engagement and recognizes that effective youth safety measures must go well beyond restrictive policies and supervisory approaches. Regulators therefore can set guidance on controllable determinants of youth online safety and work toward key pillars of digital engagement via six essential components:

Safety and Protection from Harms

Children must be protected from online harms, including harassment, exploitation, and abuse

Autonomy and Choice

Children should be respected, heard, and empowered to make informed choices

Free Expression & Information Exchange

Children should be able to freely express themselves and participate meaningfully online

Evidence-based Practices

Children need research-informed and data-driven products, policies, and protections to serve and support them

Security and Privacy

Children’s data and online activities must be protected through robust safeguards

Transparency

Children and caregivers deserve clear information about how platforms affect their rights

Recommendations for Regulators to Support Harmonized Youth Online Safety Legislation

To effectively protect youth in the digital age, EU regulators must take a comprehensive, multi-faceted approach to oversee and guide digital platforms. This includes mandating device-level age verification systems that balance security with privacy, providing clear guidance on age assurance technologies and content rating systems similar to those used in gaming and film industries, and creating incentive structures that encourage platforms to exceed minimum safety standards. The regulatory framework should also establish conformity assessments and safety standards comparable to those in the toy industry, while strengthening law enforcement efforts to address crime online. Additionally, regulators should implement robust remediation plans when safety issues are identified, similar to how other regulatory bodies like the Federal Aviation Administration operate. In the sections below, we expand on these recommendations to establish a collaborative rather than antagonistic process between regulators and platforms, ensuring that youth safety online remains the paramount priority and that platforms have clear, actionable guidance for implementing key protective measures effectively.

Mandate Verification of Age at the Device Level

*(**S**afety and Protection from Harms, **S**ecurity and Privacy)*

Historically, there has been a tension between safety and privacy on the Internet, even though at times these rights may reinforce each other. Companies need to collect and store user data to optimally safeguard users online through, for example, authentication systems, data backups, and intrusion detection systems. This can sometimes undermine privacy goals and render users vulnerable to various forms of online victimization such as identity theft, data breaches, unauthorized access to personal information, targeted phishing attacks, and potential data misuse by third parties. Additionally, this stored data could be subject to government surveillance or subpoenas, which compromises user privacy. Security may come at the cost of complete privacy when considering the intricacies and nuances of ensuring youth safety on social media platforms.

In this vein, collecting user data related to age is necessary to provide customized content recommendations, facilitate relevant advertising, and inform any age-gating approaches (e.g., parental controls, restricting mature content, constraining direct messaging capabilities and search functionality, providing time limits, and more). Unfortunately, age is not the only data point necessary to collect because minors may provide false information, use borrowed credentials, or employ unethical verification tactics to circumvent age-gating. Compounding this situation is another tension between age verification and industry-wide principles of data minimisation. Companies cannot provide a broad suite of safety restrictions and features to youth and their parents, though, without verifying their identity, their relationship, and their behaviours on the platform. As tends to be the case, the challenge lies in finding the right balance.

Historically, age verification on social media platforms (if it occurred at all) has been facilitated through the use of government ID verification, parental verification, selfies and age inference tools[134, 135] and can be used to identify which users are minors and thereby require additional protections or guardrails. However, we advocate for age verification to be implemented at the device or operating system level, and strongly recommend that regulators establish this as a mandatory standard. Unlike the fragmented approach of app-specific verification, system-level age verification provides a more streamlined, private, and uniform solution.

The key argument is that the initial point of age verification should occur when a parent or guardian first purchases and sets up a phone for their child. At this opportune moment when registering the device and signing up for cellular services, the child's birthdate can be collected and securely stored in a protected area of the device. This secure storage would be in a dedicated, encrypted section of the device's storage, often referred to as a “secure enclave” or “trusted execution environment.” These are hardware-backed secure storage areas designed to protect sensitive information from unauthorized access. Such an approach leverages the fact that parents or guardians are typically present during the initial device setup, and would provide a more reliable and trustworthy source for age information. Once this birthdate is securely stored in the device, it can serve as a centralized, authoritative source for age verification across all apps and services. This method would eliminate the need for individual apps to repeatedly ask for age information, reducing the risk of inconsistent or false information being provided.

Many child-centric non-profits including the International Centre for Missing and Exploited Children (ICMEC) and Crime Stoppers International have voiced their support for device-level age verification,[136, 137] citing the ease of implementation, enhanced accuracy and reliability, consistency and standardization, improved security and privacy, and other benefits of this approach. These organizations highlight how device-level verification is able to leverage biometric authentication and secure hardware already embedded in phones, laptops, tablets, gaming consoles, and other devices to provide a high level of accuracy that surpasses traditional methods like self-reported birthdates, which are highly susceptible to misrepresentation and circumvention.

This device-level approach places the responsibility on device manufacturers to maintain and secure this sensitive information rather than distributing it across multiple third-party companies, each with differing sets of standards, architectures, philosophies, and resources. It is also more efficient than requiring each platform to implement its own age verification process. While the sensitive nature of age verification data is undeniable, and storing it at the device level may appear to create a central point of failure in case of data breaches, this approach actually reduces overall risk compared to having multiple social media companies independently store and manage this information. By centralizing age data in a secure, encrypted format on the device, the number of potential vulnerabilities is minimized.

To reiterate, a device-level approach distributes the risk across individual devices, limiting the potential impact of a security breach to a single user’s data instead of compromising a centralized database that contains the data of all users. Age information would be cryptographically hashed and stored, with only the necessary access tokens or age-range indicators shared with individual apps. This way, platforms receive only the minimum required information to implement age-appropriate content restrictions, additional safeguards, and parental supervision tools, without having direct access to the child’s exact birthdate. This method maintains user privacy while enabling platforms to fulfill their responsibilities in protecting younger users.

While implementing age verification at the device level presents challenges, particularly in scenarios involving multiple users on a single device (e.g., siblings sharing a phone or tablet) or users transitioning between devices, these are not insurmountable obstacles. The tech industry has a proven track record of developing solutions for complex user authentication and profile management issues. For instance, modern operating systems already support multiple user profiles on shared devices, and cloud-based services enable seamless transitions between devices while maintaining user-specific settings and restrictions. Building upon these existing technologies, it is feasible to create a robust, device-level age verification system that can accommodate various use cases while maintaining security and privacy.

Guidance on Use of Age Assurance Technology

*(**S**afety and Protection from Harms, **S**ecurity and Privacy)*

Another key task for regulators is to provide an evaluation of various age assurance technologies. This is so that device manufacturers and app store providers are able to incorporate guidance on what is an appropriate balance of safety, privacy, security, and effectiveness when determining a risk-proportionate age assurance method to select. While organizations like the National Institute for Standards and Technology (NIST) have conducted a Face Analysis Technology Evaluation for Age Estimation & Verification (FATE AEV) – "an ongoing evaluation of software algorithms that inspect photos of a face to produce an age estimate," other methods should be explored so that a risk-proportionate approach to their use can be leveraged. In addition, regulators could mandate certain standards are adhered to when these companies use age assurance technologies. Particularly, regulators could mandate that companies follow the European standardisation bodies ETSI (European Telecommunications Standards Institute) and CEN-CENELEC (CEN - European Committee for Standardization / Comité Européen de Normalisation; CENELEC - European Committee for Electrotechnical Standardization / Comité Européen de Normalisation Électrotechnique Workshop), or other international organisations such as the Institute of Electric and Electronic Engineers (IEEE)[138] that have laid out appropriate and proportionate age assurance measures that:

- Adhere to data minimisation in order to be privacy-preserving, only collecting data that is necessary to identify the age, and age only, of a user;
- Protect the privacy of users in line with GDPR and other data protection rules and obligations;
- Are proportionate to the risk of harm arising from usage, and the purpose of the age assurance solution employed;
- Are easy for children to understand and considerate of their evolving capacities;
- Are secure and prevent unauthorised disclosure or safety breaches;
- Provide routes to challenge and redress if the age of a user is wrongly identified;
- Are accessible and inclusive to all users, particularly those with protected characteristics;
- Do not restrict children from services or information that they have a right to access;
- Provide sufficient and meaningful information for a user to understand how the age assurance system works, in a format and language they can easily understand – including children;
- Are effective in assuring the actual age, or age range, of a user; and
- Anticipate that users may not tell the truth, and do not rely solely on this information.

Guidance on Age-Appropriate Content Through Ratings Systems

*(**S**afety and Protection from Harms, **A**utonomy and Choice)*

While restrictions on content for children can help safeguard against seeing harmful or illegal content, regulators should provide more specific guidance on age appropriateness of content so that platforms are better able to tailor safe and relevant experiences for youth. By way of example, the video game industry's Entertainment Software Rating Board (ESRB) system demonstrates how detailed content rating frameworks can effectively guide age-appropriate experiences. The ESRB combines clear age categories with specific content descriptors that help parents understand exactly what their children might encounter in a game. For instance, when a game receives a "Teen" rating, it comes with detailed descriptors explaining whether it contains elements like fantasy violence, crude humour, or strong language. It also considers "Interactive Elements" such as unrestricted access to the Internet, in-game purchase options, "the ability to display the user's location to other users of the app" or "possible exposure to unfiltered/uncensored user-generated content, including user-to-user communications and media sharing via social media and networks" in its approach.

Similarly, in the film and movie industry, the Classification and Ratings Administration (CARA), an independent division of the Motion Picture Association (MPA) of America, provides ratings and content descriptors to help guide parents and caregivers as to whether a movie is age appropriate for their children. They also provide theatres with guardrails as to who to allow to enter a movie theatre on their own (e.g. a child would typically not be allowed to purchase a ticket and enter a theatre for adult-oriented movies in any responsibly run venue). The ratings provided as guidance are G for content that is suitable for general audiences, PG (Parental Guidance suggested) as some material may not be suitable for children, PG-13 (Parents Strongly Cautioned) as some material may be inappropriate for children under 13, R (Restricted) where a child under 17 requires an accompanying parent or adult guardian, and NC-17 where no one 17 and under is admitted. Ratings are assigned by a board of parents and guardians who consider factors such as violence, sex, language and drug use, and then assign a rating they believe the majority of American parents would provide.

In the UK, the British Board of Film Classification's (BBFC) system offers another model to consider emulating, particularly with regard to how it evaluates content impact and context. Their compliance officers are tasked with examining specific elements including dangerous behaviour, discrimination, drug use, horror, nudity, and sexual content while also considering the broader context and emotional impact on viewers. This provides an illustration of how content ratings can go beyond simple age bracketing and perceived developmental maturity to the thoughtful consideration of psychological effects and cultural sensitivities. The widespread adoption of these systems proves that structured content guidelines for youth safety can be successfully scaled. Regulators can learn from these frameworks to create standardized, age-bracketed categories for appropriateness in both content and interactive features to meaningfully guide the social media industry in the experiences they are curating for children.

The digital landscape needs comprehensive regulatory standards for age-appropriate content across social media platforms. A unified regulatory framework would establish consistent guidelines that all platforms must follow, regardless of their size or resources, ensuring uniform implementation of age-appropriate experiences. This standardization would elevate industry-wide safety measures by providing clear operational directives for content management based on specific age groups. Beyond addressing illegal content or policy violations, these standards would be particularly valuable in managing borderline content that may not warrant outright removal but requires careful age-based consideration. This includes content featuring mild nudity, mature language, substance use references, or similar material that falls into grey areas of content moderation. A tiered system differentiating appropriate content for age brackets of 13-14, 15-16, and 17 years would provide crucial clarity for platforms. To develop these standards effectively, regulators should collaborate with a multidisciplinary team of child development experts, paediatricians, mental health professionals, digital safety specialists, and parents to establish evidence-based guidelines for age-appropriate content. This collaborative approach would ensure that content standards align with developmental stages and protect young users while still maintaining engaging online experiences.

Regulators should collaborate with experts to develop comprehensive, evidence-based guidance for platforms across Europe regarding age-appropriate content and features. This guidance should address specific content suitability for narrow age brackets spanning 1-2 years, appropriate screen time limits for different developmental stages, and design features that align with young users' cognitive development. Research-based recommendations should incorporate findings on content impact, given that (for example) exposure to positive content correlates with lower depression levels while negative content shows a stronger association with increased depression.[139] The guidance should expand beyond basic content classification to encompass broader aspects affecting youth development, including mental health impacts, neurological development, self-harm prevention, and overall well-being. By establishing clear standards for content classification and protective interventions, regulators can ensure platforms implement consistent safeguards across all services accessed by youth. This standardized approach would help platforms better understand and address the varying developmental needs of different age groups while maintaining uniform protection measures across all online products and services that young people use.

Incentives for Positive Change and Driving Innovation in Safety

*(**S**afety and Protection from Harms)*

Lawmakers and regulators have incorporated substantial financial penalties into various acts and regulations, under the assumption that these fines will motivate platforms to take their responsibilities seriously and implement necessary safeguards. By way of example, the Digital Services Act (DSA) imposes significant fines for non-compliance, with penalties of up to 6% of a company's global annual turnover. Similarly, the UK's Online Safety Act (OSA) includes fines of up to 10% of global revenue or a maximum of £18 million (whichever is higher) for violations of its rules. The General Data Protection Regulation (GDPR) also carries substantial penalties, with fines of up to 4% of global annual turnover or €20 million (whichever is higher) for severe violations. The proposed AI Act includes maximum penalties of up to €35 million or 7% of worldwide annual turnover for non-compliance with prohibited AI practices.

While fines can serve as an effective deterrent, they may contribute to an adversarial rather than collaborative relationship between regulators and platforms. As a result, companies might disproportionately focus on avoiding penalties instead of innovating new safety features or otherwise improving their products and services in unique ways. To achieve a long-standing impact with minimal unnecessary obstacles, both parties should be induced to cooperate for the overall benefit of youth online. While fines have deterrent value, we believe they should be combined with incentives for positive changes. Compliance with government mandates should not only result in avoiding fines but also be part of a tiered ratings system that encourages platforms to exceed minimum compliance efforts. For instance, achieving a certain star rating or silver-level compliance could inspire a company to redouble its youth safety efforts over the subsequent year as they aim for gold-level recognition. This approach could foster healthy competition among platforms to improve baseline safety standards and drive industry-wide advancements. What is more, these regulator-provided safety ratings could be incorporated into the communications, public relations, and marketing strategies of platforms. As companies continually strive to demonstrate their commitment to user safety, they could leverage these recognitions to assure families, politicians, and other stakeholders of their achievement in meeting the highest safety standards. This positive reinforcement mechanism could then create a virtuous cycle, where platforms are motivated both by the desire to avoid penalties and to be recognised as an industry leader in online safety.

A similar model is found with the National Highway Traffic Safety Administration's New Car Assessment Program (NCAP) and Euro NCAP, which uses a 5-Star Safety Ratings program to provide consumers with information about the crash protection and rollover safety of new vehicles beyond what is required by federal law. The safety ratings cover areas such as Adult Occupant Protection (for the driver and passenger), Child Occupant Protection, Pedestrian Protection which has been expanded to include cyclists and is now known as Vulnerable Road User (VRU) protection, and Safety Assist, which evaluated driver-assistance and crash-avoidance technologies. While the rating criteria and categories of consideration would be different, this type of public rating system with standard tests and objective criteria could be a useful model to apply to social media.

Continuing with the automotive industry as a parallel for how social media platforms should approach youth online safety innovations, car manufacturers have successfully transformed safety features from mere regulatory requirements into powerful market differentiators and value creators. Recent research indicated that 42% of car buyers are willing to switch brands specifically to access superior safety technologies.[140] This clearly demonstrates how safety innovations can directly influence both consumer choice as well as market share. Given this, automobile companies compete with each other through continuous innovation in safety features like blind spot cameras, lane departure warnings, heads-up displays, and curve-adaptive headlights, social media platforms could differentiate themselves through advanced youth protection measures. Seatbelts did not give car companies a competitive advantage, nor serve as a growth driver, but sophisticated driver assistance systems do by fostering positive attention and attracting more customers. As such, platforms should adopt a similar lens and proceed in related ways. When they prioritize the safety of youth through tangible innovations and make it a core component of their brand identity and marketing strategy, their user base and engagement rates should not only grow, but also have a higher quality of experiences online.

Adherence to Safety Standards and Conformity Assessments by Notified Bodies

*(**S**afety and Protection from Harms)*

The European toy market provides a compelling regulatory model that could inform digital safety standards for children. In the EU, physical toys must meet rigorous safety requirements, demonstrated by the 'CE' (Conformité Européenne) marking, which certifies compliance with high safety, health, and environmental protection standards. Manufacturers must complete a comprehensive conformity assessment procedure, including testing, inspection, and certification, before placing products in the EU market. Such a proactive approach prevents non-compliant or unsafe products from reaching consumers. The assessment process relies on a network of notified bodies, which are organizations designated by EU member states to evaluate product conformity according to applicable legislation. These third-party assessors play a crucial role when independent verification is required by law. The UK maintains similar stringent requirements, including mandatory safety instructions and appropriate warning labels for toys.

Just as physical toys can negatively affect the health of children, so also can digital products and services provided by social media platforms. Implementing an analogous framework for these provisions could significantly enhance online safety for young users. While the specific mechanisms for enforcement would need adaptation for the online space, establishing standardized safety requirements, assessment protocols, and authorized third-party verification could create a reliable safety certification system for digital products targeting children in Europe. This approach would ensure thorough safety evaluation before they become accessible to young users, and can provide a clear safety indicator similar to the CE marking for physical toys.

Advocate for Improved Law Enforcement Efforts

*(**S**afety and Protection from Harms)*

While certain problematic online activities and behaviours are made criminal offences in some regions and contexts and not others, it is important that clarity exists for specific online harms that should be formally prohibited by law and deserving of sanction. To effectively protect youth and prevent violence online and offline, regulators must work to update criminal codes to address specific threat vectors that are historically being prosecuted using outdated and ill-fitting legislation. Further, they must advocate for, and help support, new legislation that can address novel instantiations of criminal behaviour fostered and facilitated by new technological advances. They must also demand improvements in operational protocols to support better coordinated responses, so that online misuse or abuse prompts a systematic and coordinated response, instead of one that is ad hoc, disjointed, and fragmented. Finally, they must champion better coordination of both prevention and response efforts across jurisdictions given that some of these offenses span multiple countries. If this does not happen in a timely and efficient manner, these legislative gaps will continue to provide opportunities for offenders to evade prosecution and exploit differences between legal systems as they continue to victimize other users.

The UK has significantly strengthened its approach to digital crimes through the Online Safety Act, which criminalizes cyberflashing with penalties of up to two years in prison. The legislation has elevated the sharing of intimate images without consent to a ‘priority offence,’ placing it in the highest category of online crimes alongside terrorism and drug trafficking. Under this classification, social media platforms must now implement proactive measures to prevent and remove such content or risk substantial penalties of up to 10% of their global revenue. Clear guidance on exactly what is illegal must be provided by regulators to help platforms be consistent in their implementation efforts and avoid poor judgement calls on what content or activity needs to be removed.

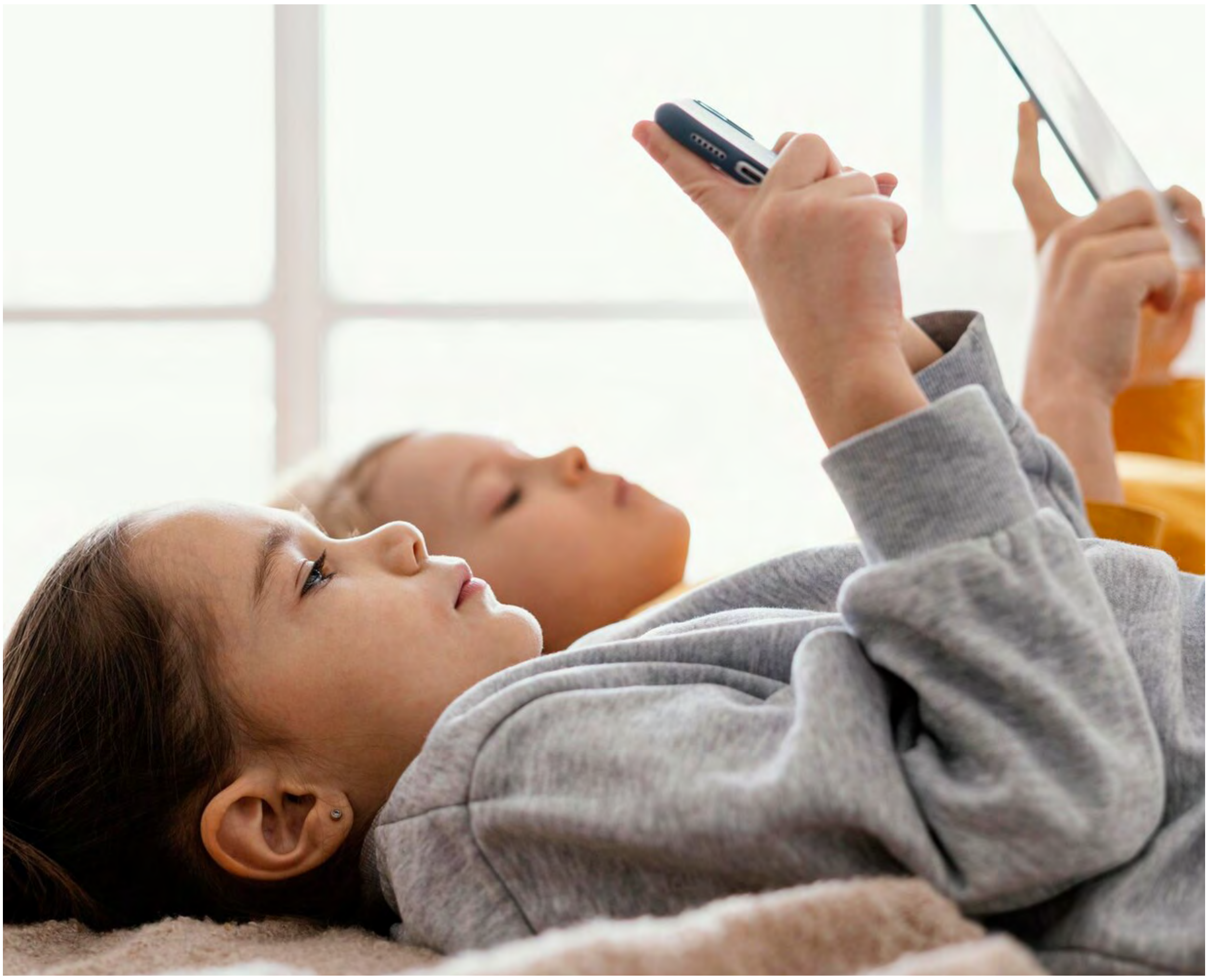
Remediation Plans

*(**S**afety and Protection from Harms, **T**ransparency)*

Regulators must have the authority to require platforms to report all discovered product safety issues for thorough assessment and enforcement action. This approach mirrors existing regulatory frameworks, such as the Federal Aviation Administration’s power to mandate safety fixes from aircraft manufacturers before planes can return to service. Similarly, Ofcom enforces strict broadcasting standards and telecommunications regulations in the UK, and can leverage its successes to develop tailored regulatory frameworks for the unique risks and operational contexts of other industries.

As another example, the United States Department of Justice (DOJ) demonstrates how regulatory bodies can effectively mandate significant changes in corporate behaviour through various mechanisms beyond traditional enforcement tools like company breakups and fines. These include implementing specific operational changes and appointing independent compliance monitors to ensure proper corporate conduct. This comprehensive approach enables the DOJ to actively shape business practices, promote ethical behaviour, and enforce robust compliance across various industries.

European Union regulators wield similar authority through frameworks such as the Corporate Sustainability Due Diligence Directive (CSDDD), which empowers them to mandate changes when companies fail to meet human rights and environmental standards. EU supervisory authorities possess broad powers, including the ability to launch investigations, impose substantial penalties of up to 5% of global turnover, and implement "naming and shaming" measures for non-compliant companies. However, regulatory effectiveness extends beyond mere enforcement powers. Regulators must provide social media platforms with clear guidance regarding identified gaps and present concrete remediation plans that incorporate specific practices, considering those previously reviewed and any new components that emerge over time. Leaving platforms to interpret vague regulations independently risks incomplete, inconsistent, or ineffective implementation of safety measures. Such ambiguity could lead platforms to either adopt a minimalist approach to compliance or implement overly broad content moderation policies that potentially infringe on fundamental rights, including children's rights as protected under the UNCRC.



Conclusion

In the EU, a harmonized framework is needed to guide the efforts of regulators and platforms with clarity and specificity as they endeavour to support youth safety and well-being online. Currently, the legislative landscape is limited by an oversimplification of the issues at hand, an incomplete understanding of how youth benefit from the affordances of social media, and a fragmented and disjointed set of directives that may not achieve the desired outcomes. To be effective, such an approach must balance the objective of comprehensive protection with the constraints of practical implementation. It should require system-level age verification (at the device or OS/App store level), standardized content classification protocols, and rigorous safety assessments before platforms can serve young users. Centralizing age verification at the device or OS/App Store level would eliminate the need for individual apps to repeatedly ask for age information in an inconsistent manner with technologies of varying accuracy and privacy. It would also reduce the risk of inconsistent or false information being provided. This approach is likely to drive improved accuracy, reliability, and privacy when it comes to developing age-appropriate online experiences, and is foundational to the broader recommendations proposed as part of the SAFEST framework.

Successful regulation in this space demands a nuanced, risk-based approach that recognizes the diverse landscape of platforms and their varying impact on young users. A streamlined approach must also take into account what is known about adolescent development and technology use, and operate in keeping with evidence-based knowledge derived from the extant research base. Just as physical product safety relies on established standards and verification processes, the harmonized framework also should provide clear benchmarks for youth protection, and build in appropriate enforcement mechanisms and regular safety audits. In addition, the framework must be outcome-focused rather than prescriptive, allowing platforms flexibility in how they achieve certain safety objectives while maintaining consistent standards across the industry. It also should embrace an iterative process that enables real-time adjustments based on user feedback and emerging risks. Platforms must demonstrate that they understand and are working to address risks proportionate to their scale and user base, and are implementing targeted interventions that match the severity of potential harms. It is incumbent upon EU regulators to establish clear guidelines while fostering innovation in online safety measures, so as to ensure that platforms not only meet minimum requirements but are incentivized to exceed them in protecting young users.

Finally, and as referenced throughout this report, we recognize that EU regulators and platforms must cooperate in ways that achieve the end goal of optimal youth safety online. The SAFEST framework and the regulator-specific recommendations provided above do not discount the critical need for platforms (regardless of size) to do their part in implementing specific research-informed policies, protocols, product features, and in-house and external initiatives to safeguard and support their user base (see Appendix B for an extensive list). Together, EU regulators and platforms must collaborate to create a digital ecosystem that not only protects youth from risks and harms, but also cultivates a pathway towards personal and professional success. This must be the incontrovertible standard for responsible innovation, both now and in the years to come.

Declaration

This project received funding support from Meta. All analysis and research were conducted independently, and all findings and conclusions are solely those of the authors.

Appendices

Appendix A

CO:RE	**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)	Identity theft, fraud, phishing, scams, gambling, blackmail, security risks
Sexual	Pornography (legal and illegal), sexuality of culture, body image norms	Sexual harassment, sexual grooming, generation and sharing of child sexual abuse material	Sexual harassment, non-consensual sexual messages, sexual pressure	Sextortion, trafficking for purposes of sexual exploitation, streaming child sexual abuse
Values	Age-inappropriate user-generated or marketing content, misinformation, 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 Risks (interpersonal, institutional, and commercial) Advanced Technology Risks (e.g., AI, IoT, Predictive Analysis, Biometrics) Risks to Health and Well-Being Inequalities and Discrimination			

Source: Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying Online Risk to Children. (CO:RE Short Report Series on Key Topics). Hamburg: Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence. https://doi.org/10.21241/ssoar.71817.

Appendix B

Below, we outline essential steps that social media platforms must take to demonstrate their dedication to youth safety—not as a peripheral consideration, but as a core element woven into their operational DNA. To be sure, many VLOPs and VLOSEs already implement many of these best practices, but we have endeavoured to serve the field by presenting a relatively comprehensive set of research-informed safety implementations to guide youth-serving platforms of varying size, developmental stage, and resource allocation. Through these steps and over time, companies should be able to demonstrate that they prioritize youth safety and well-being throughout every aspect of their products, services, and procedures. As they document and communicate their youth safety measures to EU regulators, they must convey current compliance by the requirements, and a keen commitment to continuous improvement and proactive risk and harm mitigation. All platforms should interpret these steps as an opportunity to showcase their dedication to supporting youth, rather than as a bureaucratic burden.

Utilize a Consistent Typology of Online Harms Across Industry and For Different Age Levels

*(**S**afety and Protection from Harms, **E**vidence-Based Practices, **S**ecurity & Privacy)*

There should be industry consensus around specific risks and harms that platforms must endeavour to shield youth from. One method of organization is the Typology of Online Harms published by the World Economic Forum in August 2023.[141] It categorizes various risks and threats in online spaces with a focus on six main areas: threats to personal and community safety, harm to health and well-being, hate and discrimination, violation of dignity, invasion of privacy, and deception and manipulation.[141] Furthermore, it encompasses 3 of the 4/5Cs (content, contact, and conduct risks), and acknowledges that risks may overlap and intersect and that we need to consider the rights of a child when addressing them.

Social media companies should interpret and use this typology as a common language and shared understanding of online safety risks to protect and support youth on their platforms. They should implement robust content moderation systems that can identify and address threats across all six categories, and develop age-appropriate safety features and educational resources based on this typology. Regulation should mandate that platforms view online risks and harms through the lens of this typology and be able to articulate and prove that they have aligned their safety policies and interventions accordingly. In this way, there can be a more consistent and effective approach to youth safety across the entire industry.

The Typology of Online Harms must be adapted for minors of different age brackets by considering their developmental stages and unique vulnerabilities. For younger children, the focus should be on protecting them from immediate safety risks, such as CSAM and grooming, while also shielding them from mature content that could harm their mental health and well-being. As children enter their pre-teen years, the emphasis must shift to prioritize cyberbullying, hate speech, and privacy violations, which become increasingly relevant as they begin to engage and explore more independently online. For teenagers, the typology must be applied in a comprehensive manner that addresses all six categories of harm. Given their specific vulnerabilities and developmental needs, this age group may require particular attention to harms related to romantic interactions, deception, and manipulation, as well as protection from content that may lower their self-esteem, levels of hope, perception of oneself, and general mental health.

In addition, platforms can utilize the typology to deliver age-appropriate, developmentally tailored educational content and resources across all user segments. This can include targeted push notifications, concise instructional videos created by the platform or influencers, and timely reminders about available in-app safety controls. These efforts should foster mature decision-making, enhance digital literacy skills, and promote personal agency in online actions and interactions. Of course, platforms also must remain mindful of teens' growing need for personal privacy, access, and greater independence.

Striking the right balance is crucial to ensure that protective measures are neither overly restrictive for older minors nor too lenient for younger ones. Companies must remain agile, adapting their approaches to accommodate industry-wide changes, regulatory recommendations, and the naturally evolving capabilities of children over time. Said another way, the safety approach of platforms must grow with the child. Legislative approaches to youth online safety must encourage and empower platforms to continuously refine their safety measures to align with the most current industry-wide, academic, societal understanding of online risks and child development, and do so in a way that does not create an undue burden.

Social Privacy Settings as Default for Youth

*(**A**utonomy and Choice, **F**ree Expression and Information Exchange, **S**ecurity and Privacy)*

As reviewed earlier, EU legislative efforts to promote online safety have emphasized the need for age-appropriate experiences, transparency, improved reporting, and heightened protections from harmful content. We also mentioned that the UK's Age-Appropriate Design Code (Children's Code) is the only piece of legislation in the UK or EU that explicitly requires "high" privacy settings by default, while in the US, KOSPA requires the "highest" privacy settings by default. In September 2024, Instagram launched what they termed "Teen Accounts," where accounts are automatically set to Private for all users under 18. This means that only approved followers can see their content and interact with them. Users who are 16 and 17 can adjust this setting to open themselves up to broader interactions and visibility, but users who are 13 to 15 can only do so with parental permission through Instagram's Parental Supervision tool. This was a unique and notable development in the industry; we expect other platforms to follow suit in the near term.

Blocking, Muting, Filtering, and Reporting

*(**S**afety and Protection from Harms, **F**ree Expression and Information Exchange)*

To address contact risks, almost every platform with interactive functionality provides tools to enable users to report, block, and mute others.[142-145] Apart from these manual controls, developments in AI are helping prevent contact risks. For example, Reddit released a new Harassment filter in the Spring of 2024 powered by a new LLM trained on flagged content and moderator actions. This can be used in conjunction with their Mature Content filter, which uses automation to filter out content that is sexual or violent. They also improved their reporting capacity so that individuals can report if specific components of another user's profile violate platform policy. These new iterations upon existing safety features are welcomed, and further revisions will be important to provide Reddit users with positive experiences continually.

> **“Reporting should be easier, but I get why it’s a double-edged sword. It could lead to false reports. I’ve seen things online that don’t affect me personally, but should be reported. The problem is, when I go to the reporting section, they make me click through three subsections, add a comment, and describe exactly why from 0 to 10. At that point, I just cancel and move on because it’s too much hassle.”**

— Male, 16 years old, Spain.
ThinkYoung Focus Group 2,
12th of November, 2024.

Enhanced Parental Controls

*(**S**afety and Protection from Harms, **S**ecurity and Privacy, **A**utonomy and Choice, **F**ree Expression and Information Exchange)*

The DSA and KOSPA also mandate the provision of better parental controls, while the OSA calls for new requirements regarding the terms of service, which need to detail how children are to be prevented from encountering primary priority content that is harmful to them. Social media platforms have strengthened their safety measures in response to this evolving regulatory landscape. For example, Instagram's "Teen Accounts" implementation launched in September 2024 represents a significant shift in youth protection. The platform now requires Parental Supervision for users aged 13 to 15 who wish to modify their privacy settings or request extended screen time beyond parent-set limits. This marks a departure from the previous opt-in model, where both parent and child could choose whether to enable supervision features. The new system makes parental oversight and involvement mandatory for a child to make certain account modifications, and is hoped to facilitate more communication and collaboration between parents and children when it comes to healthy and positive social media usage.

Instagram's Teen Accounts provides numerous additional features to support child safety and well-being objectives. Parents can monitor which accounts their child has DMed in the last seven days (without seeing the contents of any DMs, which are protected for privacy reasons), restrict access to the app during certain periods of the day or night, and view the topics their teen has chosen to follow. In addition, messaging and tagging will be restricted; teens under 18 can only receive DMs from, or be tagged or mentioned by, people they already follow. Additionally, time management features have been provided; for example, Instagram will send reminders to teens under 18 after one hour on the platform, in an effort to encourage them to go do something else. A "sleep mode" has been implemented to silence notifications and send automated replies to direct messages between 10 p.m. and 7 a.m. (which can be adjusted with parental permission) to encourage healthier bedtime habits. Parents will also be notified when their teen blocks or reports someone. Finally, teens under 18 will be placed under Instagram's most rigorous content control measures, which means that they will be protected from [what Meta has determined as] "sensitive" or mature content on their Explore page or in Reels. Relatedly, they will be under the most restrictive version of the "Hidden Words" anti-bullying feature, so offensive words and phrases will be filtered out of the Instagram comments and DM requests received.

It is arguable that these significant changes were prompted in part by EU and US legislative developments. In the summer of 2024, other social media platforms rolled out similar, though smaller-scale, protective measures to more optimally protect and support youth on their platform. For instance, Snapchat implemented new measures in June 2024 to protect teens from potential exploitation. The platform now displays enhanced warning notifications when teens receive messages from users outside their mutual friends list or from blocked individuals. The system also alerts teens about messages from users in regions associated with scam activities. Additionally, Snapchat strengthened its blocking tools to prevent circumvention through new accounts and increased the frequency of location-sharing setting reminders for its Snap Map feature. In September 2024, YouTube launched its new supervised experience that enables parents to link their accounts with their teens'. This system provides notifications when teens upload videos or start live streams, allowing parents to monitor subscriptions, comments, and general platform activity. The feature creates a collaborative approach to supervision, where both teens and parents maintain mutual control over the experience, with either party able to deactivate supervision if needed. Finally, TikTok offered an initial version of its Family Pairing system in 2020, but it has been updated periodically to provide more features and functionality. It allows parents to link their accounts with their teens' accounts to set daily screen time limits, enable "Restricted Mode" to filter out potentially mature videos, control direct messaging permissions, and limit searchable content.

Formal Family Onboarding Processes

*(**S**ecurity and Privacy, **T**ransparency)*

Historically, societies have marked the transition from childhood to adulthood with various rites of passage, signifying increased responsibilities and privileges. In the digital age, allowing teens to use social media platforms has become a modern rite of passage that requires careful guidance and preparation. Social media companies need to recognize this transition and create a formal "onboarding process" for both parents and teens, mirroring the structured approach of traditional coming-of-age rituals. This on-boarding process should be comprehensive, interactive, and tailored to both teens and parents, ideally through the engaging and convenient medium of short-form video. For teens, it should cover essential aspects of platform use, including privacy settings, safety features, reporting mechanisms, responsible use, critical thinking skills, and digital literacy skills. For parents, the process should provide clear guidance on how to use parental controls, monitor activity, and engage in open conversations about online behaviour. Scenario-based exercises with questions to answer and considerations to ponder should also be a part.

By framing this process as a digital rite of passage, social media companies can emphasize the importance of trust, responsibility, and maturity in online interactions. This approach enhances safety and promotes a healthier relationship between teens, parents, and technology, potentially reducing conflicts and misunderstandings about social media use. Just as traditional rites of passage prepare youth for adult roles in society, a digital onboarding process can help teens navigate the complexities and nuances of online spaces while ideally maintaining the trust of their parents and guardians. Moreover, this process should actively encourage, empower, and equip parents to be involved in their children's online lives, providing them with the knowledge and tools to proactively safeguard their teens' digital experiences. By fostering open communication and shared understanding between parents and teens about online safety and mature decision-making, this onboarding process can facilitate a supportive framework for healthy and appropriate digital engagement, ensuring that parents remain informed and involved partners in their children's online journey.

Youth-Specific Controls and Improvements

*(**S**afety and Protection from Harms, **E**vidence-based Practices, **S**ecurity and Privacy)*

As yet another improvement, Instagram implemented a "Nudity Protection in DMs" feature in April 2024 using on-device machine learning and without requiring Meta to have access to users' nude images unless they are reported. Enabled by default for users under 18, this feature automatically blurs images detected as containing nudity so that the recipient is not immediately confronted with it. They also receive a message encouraging them not to feel pressured to respond, and are provided options to block and report as they see fit. For those who send nudes, they receive a message that reminds them to be careful sending sensitive images, and an option to unsend them.

With regard to sextortion prevention, Instagram has developed algorithms to identify potential sextortion accounts based on behavioural patterns and prevents these accounts from viewing teens' profiles in follower lists or search results. Additionally, Instagram has implemented screenshot and screen recording prevention where recipients of a photo or video in a private Instagram message created with the "view once" or "allow replay" feature are not allowed to screenshot or screen record it without the sender's consent. Instagram also hides the "Message" button on teenagers' profiles from potential sextortion accounts (identified through analyses of behaviour, location, and historical activity) in an attempt to deter continued communication.

As a final example, Snapchat employs sophisticated algorithmic systems to protect users from potentially harmful interactions. The platform's safety architecture, enhanced in June 2024, uses machine learning to detect suspicious behavioural patterns and automatically flags problematic accounts. When these accounts attempt to communicate with young users, the system displays prominent warning messages and safety recommendations. The platform's anti-circumvention technology extends beyond simple blocking by implementing device-level restrictions by preventing banned users from creating new accounts on previously blocked devices. Moreover, Snapchat's geographic risk assessment system automatically restricts friend requests from accounts in regions associated with high rates of scamming activity, particularly when these accounts lack mutual connections with the target. This multi-layered approach combines behavioural analysis, device fingerprinting, and location-based risk assessment to safeguard youth on their platform.

The recent surge in youth safety innovations across social media platforms reflects both legislative pressure and growing societal awareness of online risks. The aforementioned regulatory efforts have catalysed significant technological advancements in youth protection, ones that are needed to keep pace with current and emerging risks to youth safety and well-being. Given the dynamic nature of online risks and the constant need for evolution in protective measures, ongoing investment in research and development of safety technologies remains paramount. Continued proactive innovation that meets the current needs of teens and families in the form of easy-to-use, practical tools is essential to demonstrating platforms' genuine commitment to youth safety beyond simply regulatory compliance requirements.

Avoidance of Secondary Victimization

*(**S**afety and Protection from Harms, **F**ree Expression and Information Exchange, **S**ecurity and Privacy)*

From the field of criminal justice stems the concept of secondary victimization, defined as "negative social or societal reaction in consequence of the primary victimization and is experienced as further violation of legitimate rights or entitlements by the victim".[146] Said another way, "following the loss of control that often accompanies criminal victimization, victims seek recognition and support, and professional but distant reactions from authorities can leave victims feeling rejected and not supported".[147]

Generally speaking, if a victim of interpersonal harm has a poor experience with authorities who are supposed to respond to their call for help, they feel doubly victimized. Research is clear that victims often feel re-violated due to the insensitive or inadequate response of those who are supposed to come to their aid, such as when they fail to recognize the gravity of the offense or display empathy toward the victim's experience.[148, 149] Incomplete follow-up or infrequent contact with the victim can also produce high levels of uncertainty and a deep lack of trust, which can result in the victim choosing not to report any future incidents.[147, 150]

For some victims, being treated in this way may actually be more harmful than the original victimization.[151] and can lead to various forms of felt trauma [152] Secondary victimization has been correlated with posttraumatic stress symptoms and physical and psychological distress.[147, 149, 153] In addition, the target's self-esteem, faith, and trust in the system – and society at large – may very well be compromised permanently.[151, 152]

However, research is also clear that having positive interactions with authorities in charge of responding is incredibly important for the victim's recovery process.[147] Remaining "in the know" and feeling supported by caring, conscientious responders can reduce depressive symptoms and enhance quality of life, while also assisting in healing and rebuilding their lives.[151] Indeed, the manner in which the victim is treated throughout the process, the amount of control the victim is given, and the extent to which they are allowed to participate all greatly influence the victim's mental and physical well-being. Communication is the key; it helps victims feel they are involved, know what to expect with the investigation and adjudication of the matter, and take comfort in receiving regular updates – which is tied to their feelings of safety.[147, 151]

While some social media platforms have used AI and automated methods to provide immediate acknowledgement and a subsequent update to those who file reports after being targeted or harms, others respond in an incomplete manner, or not at all. We acknowledge it is extremely difficult to keep up with the volume of inbound reports, and challenging to interpret what is submitted when violations are unclear, context is missing, cultural differences are implicates, and/or screenshots and screen recordings were not included. Nonetheless, technological solutions in this regard that can provide systematic, prompt, and regular updates to those who report must be constructed and implemented. Social media companies must – at all costs – keep their users from being victimized a second time because they failed to respond to a report of abuse or harm. We assert that a 24-hour initial response time for platforms to address user reports of harm should be mandated. This does not mean the issue is resolved, but that the report has been received and has triggered progress through the verification, investigation, and response workflow. While a 24-hour required initial response time may be challenging and seems only aspirational, we believe it is eminently achievable. Moreover, the speed at which takedowns and formal responses occur is of great importance when considering the need to combat misinformation, disinformation, and the viral spread of harmful interpersonal content such as targeted harassment and deepfakes. Given the uneven experiences of social media users after filing reports, such legislative mandates in this area seem appropriate and due.

Age / Maturity Filters

*(**S**afety and Protection from Harms, **A**utonomy and Choice, **S**ecurity and Privacy)*

As yet another example to forestall content risks, platforms can use labelling and filtering systems to guide users about the nature and suitability of the content, such as age ratings, parental controls, or sensitivity screens. These systems can help to empower parents and guardians to make informed and responsible choices about the content that they or their children can see. Facebook, Instagram, X, and TikTok provide word-based filtering tools to implement such constraints.[154] TikTok's content filtering system allows users to customize their viewing experience through a simple keyword blocking feature. Located in Settings and Privacy under Content Preferences, the "Video Keywords" filter prevents specific content from appearing in both the "For You" and "Following" feeds. When keywords are added to this filter, TikTok automatically blocks any videos containing these terms in their descriptions or stickers, giving users (and parents) greater control over the content they encounter. Similarly, Instagram's Sensitive Content Control, Snapchat's Restrict Sensitive Content, and TikTok's Restricted Mode allow users to control the amount of content they see that may be upsetting, offensive, sensitive, or suggestive in their feeds, recommendations, or search results.

Enhanced Content Moderation

*(**S**afety and Protection from Harms, **S**ecurity and Privacy)*

Content moderation should be a foundational aspect of platforms' efforts to address child safety. This involves enforcing established policies by removing content or taking any other action that is in line with the rules set forth. According to research published by the International Centre for Missing and Exploited Children (ICMEC), content moderation is the review of user-generated content against platform policies, conducted through a combination of human and technology review. Content moderation helps protect children online by removing or blocking harmful content, such as cyberbullying, hate speech, and inappropriate or sexually explicit material, from online platforms. By doing so, content moderation creates a safer online environment for children by reducing children's exposure to harmful content and behaviour. This can also help prevent re-victimization: by quickly removing known violating content, platforms prevent continued harm to children that is present when content continues to circulate across multiple platforms. To deep dive into the right content moderation practices for a given platform, it is recommended that platforms consult further guidance by ICMEC (including a recent Model Framework for Employers of Content Moderators), the TSPA, and other such trusted organizations. Specific decisions around use of automation, outsourcing, and other aspects of content moderation operations should be made to minimize risk to children.

Fact-Checking, Warning Labels, and Watermarks

*(**S**afety and Protection from harms, **A**utonomy and Choice, **E**vidence-Based Practices)*

Social media companies can also use fact-checking and warning labels to flag content containing misinformation, such as false or misleading claims about COVID-19, elections, vaccines, or synthetic content (e.g., produced by generative AI). To that end, in May 2024, TikTok began to tag any AI-generated content with metadata called "Content Credentials" – a watermarking technology to help establish provenance and to inform users that what they are seeing is not organic. In preparation for the 2024 US elections, Meta applied "Made with AI" labels to AI-generated content, and more prominently highlighted digitally-altered media that poses a "particularly high risk of materially deceiving the public on a matter of importance".[155] Such an approach can help to reduce the exposure and impact of harmful or inappropriate content on all users (including youth), and to enhance the quality and credibility of the information online.[156, 157] Platforms can also pursue innovative solutions such as "Community Notes" (on X), where veteran contributors who have no previous violations can add helpful context to potentially misleading posts. These notes then become visible only after reaching a critical threshold of "helpful" ratings from other contributors, and can thereby provide valuable information to others as they interpret the truthfulness of posts. Crowdsourced fact-checking will continue to evolve and should be considered more often as a method to combat misinformation.

Partnerships with Third Party Researchers

*(**A**utonomy and Choice, **E**vidence-Based Practices, **T**ransparency)*

Learning from findings from nationally representative research on youth experiences online is critical for both social media platforms and society as a whole. This type of research provides valuable insights into the evolving digital landscape and its impact on youth, and helps platforms understand trends, potential risks, and opportunities for education, interventions, feature development, and other enhancements. To ensure objectivity and avoid conflicts of interest, as well as the optics of potential bias and influence, it is essential that this research be conducted by independent third-party researchers rather than the platforms themselves. This approach not only enhances credibility but also allows for a platform-agnostic view of youth experiences while avoiding corporate blind spots, selection bias from convenience sampling and anecdotal accounts, and a lack of appropriate representation. Regular data collection, conducted at least annually, enables the tracking of longitudinal trends and the discovery of novel insights related to the prevalence, correlates, and causes of internet attitudes and behaviours among youth.

As one idea, a formalized joint research fund housed within a respected academic institution should be set up to create a sustainable ecosystem for rigorous, independent research on youth online safety. This fund would operate through mandatory contributions from platforms, and its governance would be overseen by an independent review committee comprising academic experts and youth advocates while also supported by an advisory board representing diverse stakeholders from academia, industry, youth organizations, and legal sectors. This structure ensures both scientific rigor and practical relevance while maintaining independence from platform influence. Through open calls for proposals, the fund would support research grants of varying amounts, with clear milestone requirements and accountability measures.

Research priorities would be established through collaborative dialogue between regulatory bodies and platforms, with input from academic experts, youth advocates, and child safety organizations. This approach ensures that funded research addresses both emerging regulatory concerns and practical platform challenges while maintaining focus on youth mental health and well-being as the paramount consideration. Annual priority-setting meetings would allow for timely adjustments to research focus areas based on evolving digital risks, technological developments, and observed patterns in online behaviour. Grant recipients would be required to submit regular interim and final reports, participate in stakeholder workshops, and publicly disseminate their findings. By supporting such independent studies, platforms would demonstrate their commitment to data-driven knowledge, transparency, and youth online safety while gaining insights to refine their products, services, policies, and programming.

Independent Platform Advisory Boards

*(**A**utonomy and Choice, **E**vidence-Based Practices)*

Many of the social media platforms have established expert advisory boards comprising professionals from diverse fields to solicit research- and practice-informed guidance on how best to support youth within the products and services they build. For instance, Meta’s Oversight Board consists of experts who provide quasi-judicial review of content moderation decisions, while Snap’s Safety Advisory Board includes professionals from various disciplines including online safety organizations, academia, and mental health. TikTok’s Content Advisory Council brings together distinguished experts in technology, policy, and health and wellness to inform their content moderation policies. YouTube’s

Child Safety Advisory Committee plays a key role in shaping the platform's services that affect young users, especially as it relates to age-appropriate experiences. When properly empowered, these expert councils enhance platform legitimacy and safety practices by contributing evidence-based guidance, challenging problematic policies, and pushing for greater transparency in platform operations. While what platforms ultimately do remain their sole discretion, the input of advisory boards attempts to ensure that the trust and safety measures implemented are grounded in empirical research and child-focused best practices.

Researcher Access to Anonymized User Data

*(**A**utonomy and Choice, **E**vidence-Based Practices)*

Related to this, social media companies historically have been reluctant to provide researchers with access to user data, citing privacy concerns and potential misuse. This aversion was notably intensified following incidents like the Cambridge Analytica scandal in 2018, which led to widespread criticism of Facebook's data-sharing practices and resulted in a $5 billion FTC fine. As a result, companies like Meta (formerly Facebook) and X (formally Twitter) significantly restricted access to their APIs and user data for research purposes, limiting the ability of academics to study online behaviours and platform effects.

However, there are signs that this stance may be shifting, as some companies are exploring ways to collaborate with researchers while still protecting the privacy of the data. For example, in Fall 2024, Meta partnered with the Center for Open Science (COS) on a pilot program to share certain Instagram data with select academic researchers. This initiative aims to facilitate studies on the social and emotional health of teens and young adults in a privacy-conscious manner. The program employs a "Registered Reports" model, where peer review is conducted before data collection and analysis, which is intended to promote transparency and reduce potential bias.

While this represents a step towards greater openness, it is important to view such initiatives objectively. The pilot program is limited in scope, and it remains to be seen how extensively it will be implemented or expanded. Moreover, the data shared is still controlled and filtered by Meta, potentially limiting the breadth of research questions that can be addressed through the constrained data access provided. However, this endeavour should be applauded and is potentially groundbreaking. It is critical for platforms to support this type of research. By helping to facilitate independent studies, they demonstrate a commitment to external partnerships and collaborative efforts to solve challenging problems, and implicitly agree to increased transparency and accountability. More importantly, the insights gained from such research can be incredibly valuable for platforms in refining their products, services, policies, and programming.

Assessing Risks through Child Rights Impact Assessments (CRIAs)

*(**S**afety and Protection from Harms, **F**ree Expression and Information Exchange, **E**vidence-Based Practices)*

The UN Guiding Principles on Business and Human Rights (UNGPs) and the Children's Rights and Business Principles (CRBPs) establish a clear responsibility for all companies, including social media platforms, to identify and mitigate any adverse human rights impacts associated with their operations. This is especially important when it comes to children's rights given their unique vulnerabilities and the long-term implications that violations of their rights can have on their development and future. Under the UNGPs and CRBPs, a Child Rights Impact Assessment (CRIA) serves as an essential instrument to analyse the effectiveness of their current products, services, policies, and procedures in addressing various child rights issues as defined in the UNCRC. CRIAs are not currently mandated by law, but they should be. By conducting a CRIA, platforms can gauge the extent to which each identified child rights concern is adequately managed within their existing processes.

Furthermore, this assessment facilitates the development of a more integrated and holistic strategy for safeguarding child rights throughout the organization's structure and activities. This is especially important because research finds that most companies perform risk-based assessments but not rights-based assessments. Ultimately, a CRIA not only measures and highlights areas of concern but also provides a foundation for implementing robust child protection initiatives while also reinforcing a corporate ethos of responsibility towards children's well-being across all organizational contexts.

At the start of a CRIA, platforms must first understand the risks, harms, interventions, laws, and research at the intersection of youth online safety and child rights. The authors of this report have attempted to provide that backdrop, but it must be acknowledged that the digital landscape, legislative backdrop, and the context of youth development are continually changing. As such, care must be taken to stay connected and informed with the latest developments in these areas. Platforms must also utilize youth councils (explained below) and updated internal and external data in order to promote stakeholder engagement and properly canter youth in their assessment initiatives. This will ensure that the assessment items align with the lived experiences of young people on their platforms and will likely uncover previously neglected or unknown areas which merit inquiry, attention, and response.

Major social media companies collect and analyse a vast array of internal data about their users, which can be leveraged to inform their efforts to safeguard and support youth online. These data points typically include profile information provided upon account creation, behaviour metrics like time spent on the platform, content consumption patterns, engagement metrics, and usage of various platform features. Social graph data can also provide insight into users' social networks and relationships, including the number and demographics of connections, interaction frequency, and participation in various group chats or channels. Additionally, platforms have data points related to safety-related actions by users, such as blocking, muting, reporting content or users, and privacy setting choices. Analysing these data should provide insights into the types and rates of harms reported to the platform over time, and how they are distributed across different demographic groupings. This analysis can help identify trends in online risks and challenges faced by youth users, allowing platforms to develop more targeted and effective safety measures. The data can also shed light on the usage rates of existing safety controls offered by the platform, which may suggest areas for improvement or the development of new safety features.

All of this said, it is important to recognize that internal data alone offers an incomplete picture, as many victimizations and risks are underreported by users to the platform.[158-160] Therefore, it is essential to complement this with external data derived from quantitative and qualitative findings by objective third-party researchers conducting nationally representative research on youth experiences on the platform. These studies should employ a variety of valid and reliable scales developed by social media scholars, as documented in the existing literature. As research in this field evolves, new and improved measurement instruments specific to online safety issues continue to be developed and refined. All research activities must adhere to ethical standards, including voluntary participation, obtaining parental consent and child assent, ensuring confidentiality of internal data, and maintaining anonymity of external data. Additionally, the language used in these research initiatives must be developmentally appropriate, culturally specific, and sensitive to the diverse backgrounds of participants.

Finally, platform trust and safety personnel must review all relevant findings to determine which risks and harms are most significant in severity, scope of impact across the user base, potential for and velocity of growth or virality, frequency of occurrence, and potential for long-term psychological or social consequences. The severity of risks and harms should be assessed based on their potential impact on individual users and the broader community. The scope of impact across the user base helps identify issues affecting many users or specific vulnerable groups. The potential for and velocity of growth or virality is crucial in understanding how quickly a risk or harm can spread and escalate on the platform. Assessing the frequency of occurrence helps trust and safety personnel allocate resources accordingly. The potential for long-term psychological or social consequences is a critical factor, considering that youth professionals are increasingly examining interpersonal harm (offline and online) through a trauma-informed lens. That is, they are rightfully viewing and taking into account the lasting negative effects of online risks and harms on young users.[161]

Transparency Reports

*(**E**vidence-Based Practices, **T**ransparency)*

Social media platforms regularly publish transparency reports that provide insights into their trust and safety efforts. While some variations in the specific metrics are reported, many of the key data points overlap across platforms. Common ones include the number of policy violations, content removals, account suspensions or terminations, appeals of moderation decisions, response times for addressing reports, and the prevalence of violating content. They also often provide high-level numbers on government requests for user data, content removals due to copyright claims, and measures taken to combat spam and fake accounts. Some platforms provide more granular breakdowns of violations by category, such as hate speech, harassment, violence, or CSAM content; this should be the standard, and others should follow suit and provide similar comprehensive details.

Transparency reports, while informative, often leave significant gaps in understanding youth experiences on social media platforms. Key areas that require further elucidation include the encounters that young people have with harmful content, misinformation, and manipulated media, as well as their exposure to spam or fake accounts. There is also insufficient data regarding user awareness and utilization of reporting tools and safety features, their experiences with content moderation decisions, and their perceptions of platform responsiveness to reports. More comprehensive data is needed on the frequency with which youth encounter content that violates platform policies, their reporting behaviours, and satisfaction levels with platform responses to these reports. Additionally, understanding overall feelings of safety while using these platforms remains crucial yet understudied. To address these knowledge gaps and provide a more comprehensive view of user experiences, platforms should consider implementing surveys, focus groups, and other mixed-methods research initiatives. These approaches can offer valuable insights into the nuanced realities of platform usage and help inform more effective safety and moderation strategies.

The Ability to Reset Recommendation Algorithms

*(**A**utonomy and Choice, **T**ransparency)*

Due to increasing legislative pressure, many platforms are introducing features that allow users to reset their algorithms and start fresh. In 2022, TikTok rolled out a feature that enables users to adjust their Content Preferences and reset the algorithms influencing their For You Page (FYP). In November 2024, Instagram followed suit with a feature called “Fresh Start,” which lets users reset their Feed, Explore Page, and Reels. This functionality helps users break free from filter bubbles, echo chambers, and the rabbit holes of potentially harmful or negative content and encourages exposure to more positive, wholesome, and beneficial content. However, controlling one’s recommendation algorithms on other platforms is more difficult. YouTube offers a feature where users can clear their watch and search history, and adjust their interests within the platform, which directly influences the recommended content they see. Facebook also provides users options to modify their feeds by unfollowing accounts or pages and hiding specific posts, which indirectly influences the recommendations they receive. X (formerly Twitter) enables users to manage the list of whom they are following, change their topics of interest, and mark some content as “Not Interested,” which informs the platform’s algorithms that push content into their feed.

Recent research by Project Rockit,[162] based on a sample of over 1,000 young people, found that 56% of respondents desired the ability to reset their recommendation algorithms. At a minimum, all platforms that rely on algorithms to curate content should offer users the option to start anew. Ideally, these platforms should provide easily accessible reset options that are user-friendly and come with comprehensive walkthroughs. Users should be able to review and adjust various factors influencing algorithmic decision-making, such as followed accounts, liked or favourited content, advertisement preferences, content category preferences, interaction history, search history, and device and app usage patterns. By offering these options, platforms can give users greater control over their online experiences, potentially promoting more positive and diverse content consumption. This approach would help to enhance the user experience by ensuring that recommendation systems are aligned with users’ preferences and well-being, rather than reinforcing narrow or harmful content loops.

Data Deletion and Portability

*(**A**utonomy and Choice, **T**ransparency)*

In recent years, major digital platforms have enhanced user control over personal data by introducing features that allow for data deletion and portability. This shift has been largely influenced by global privacy regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Users can now request the removal of their data from a platform’s servers or obtain a copy of their information in a structured format. To access these features, users should navigate to the Privacy section of their app settings, where they will find options related to data deletion and portability. The specific terminology and processes may vary by platform, but the core functionalities remain consistent. These enhancements not only empower users but also ensure compliance with legal standards, fostering trust and transparency in the management of personal information.

Behavioural Nudges

*(**E**vidence-Based Practices)*

To address conduct risks, platforms employ behavioural nudges through architectural changes and targeted messages to promote prosocial behaviours and community norm compliance. In late 2019, Instagram started using a Comment Warning system to encourage users to pause, reflect, and edit their words before they shared something potentially offensive or hurtful. Relatedly, in 2020, Facebook began to let users know when their about-to-be-posted comments or captions were similar to content that had been previously reported as abusive. In early 2021, TikTok started to automatically detect language that violates their policies and allows users a chance to edit or discard their post. In 2020, YouTube reminded users to "keep comments respectful." They also began to use popups to introduce cognitive dissonance and encourage individuals to reconsider their comment before it is shared on another person's post (e.g. "Would you like to reconsider posting this?"). In 2022, Instagram reminded users to help keep its platform "a supportive place." Also in 2022, LinkedIn introduced nudges (e.g., "Please keep LinkedIn respectful and professional") to encourage positive behaviour among users who had previously posted inappropriate content. In August 2024, Facebook launched a new feature that helps creators avoid "Facebook jail" (where their abilities to post and interact are greatly limited) by allowing them to take a short educational course that reminds them about the rules related to posting inappropriate content.

Platforms also often can provide nudges about, and links to, Community Guidelines so that users are educated about behavioural expectations towards others. Relatedly, apps can pre-empt conduct risks by promoting healthier screentime habits. For example, prompt reminders on TikTok state, "Break reminders help you feel more mindful and balanced on TikTok."

All of this said, an exploratory research project in 2024 involving 4,000 US adults found that nudges within a Facebook-like news feed (called "Mock Social Media Website Tool") did not significantly decrease engagement with hate speech (sharing, commenting, or reacting to it), but did increase engagement with harmless and wholesome content.[163] Much more research is required to fully understand the value of nudges, and they may prove more effective with youthful populations, on youth-centric platforms, and/or with other forms of harm. Overall, these techniques and approaches may have promise in motivating and reinforcing users' prosocial and ethical inclinations towards others.

Initiatives to Educate Youth and Families

*(**A**utonomy and Choice, **E**vidence-Based Practices, **T**ransparency)*

Social media platforms play a crucial role in educating and safeguarding youth online, extending beyond mere content moderation to active engagement in user education. These platforms are uniquely positioned to inform users about potential risks, emerging harms, and available safety features through various innovative approaches. By developing user-friendly guides, creating short-form videos, and designing interactive resources, platforms can effectively communicate complex safety concepts to both young users and the adults who support them. These educational materials can cover a wide range of topics, from basic online safety practices to more nuanced issues like metaverse safety, parental controls, and strategies for engaging in productive conversations about digital well-being.

To maximize the impact of these educational efforts, platforms can utilize multiple dissemination channels, including dedicated safety centres on their websites, direct messaging to users, targeted social media campaigns, and strategic partnerships with schools and community organizations. Additionally, hosting and sponsoring community-wide trainings for parents and educators, either through local events or virtual platforms like Zoom, can help address emerging online phenomena related to metaverse risks or generative AI harms. Platforms can further support this educational mission by developing comprehensive toolkits for families and schools, designed to promote digital citizenship and media literacy. These resources might include interactive modules, discussion guides, and age-appropriate activities that empower young users to navigate online spaces safely and responsibly. By leveraging their technological expertise and extensive reach, social media companies can create engaging, accessible content that resonates with youth and their caregivers, ultimately fostering a culture of online safety and digital well-being. This proactive approach to education positions platforms as partners in cultivating responsible digital citizens and building healthier online communities.

Hiring and Building-out Trust and Safety Teams

*(**S**afety and Protection from Harms, **E**vidence-Based Practices, **S**ecurity and Privacy)*

AI-driven content moderation has made significant strides in recent years, leveraging natural language processing and machine learning to proactively detect and filter out a large volume of harmful content. These systems can rapidly scan vast amounts of data, identifying patterns and potential violations of community guidelines at a scale impossible for human moderators alone. AI moderation tools are particularly effective at detecting explicit content, hate speech, and known patterns of abusive behaviour, providing a critical first line of defence in protecting youth online. However, human moderation remains indispensable in addressing the nuances and contextual complexities that AI may struggle with. Human moderators bring critical thinking, cultural understanding, and empathy to the moderation process, allowing for more nuanced decision-making in ambiguous cases. They play a vital role in reviewing edge cases, handling appeals, and fine-tuning AI systems based on evolving trends and user behaviours. Additionally, human moderators can identify emerging threats and adapt strategies more quickly than automated systems alone. Social media platforms must continue to invest in building robust Trust and Safety teams, recognizing that human expertise remains key in addressing nuanced content moderation challenges.

While some smaller platforms may be tempted to outsource content moderation to third-party providers and reduce their in-house Trust and Safety staff, this approach is actually detrimental to platform safety and user experience. Trust and Safety personnel play an irreplaceable role that extends far beyond overseeing and complementing automated moderation systems. These professionals are instrumental in developing comprehensive platform policies and community standards that reflect the unique needs and values of the platform. They ensure compliance with an ever-evolving landscape of global regulations, which is particularly critical in today's complex digital environment. Trust and Safety personnel are also at the forefront of addressing sophisticated challenges such as bot activity, web scraping, misinformation and disinformation campaigns, and fraud detection. Moreover, they handle the nuanced task of managing user reports and appeals, requiring a deep understanding of context and platform-specific issues that automated systems do not have. Perhaps most importantly, these teams are responsible for devising and implementing various risk and harm mitigation strategies, anticipating potential issues before they escalate into major problems. By undervaluing the multifaceted role of Trust and Safety personnel, platforms risk compromising not only their users' safety but also their long-term sustainability and reputation especially given the level of scrutiny they will increasingly be under from regulators and other stakeholders.

Youth Safety Councils

*(**A**utonomy and Choice, **F**ree Expression and Information Exchange)*

Several major social media and gaming platforms, including Snap, TikTok, and Roblox, have established Youth Councils to engage directly with their young users and gain insights into their experiences, concerns, and ideas for improving online safety and well-being. These initiatives aim to empower youth voices in shaping platform policies, features, and safety measures. Roblox's Teen Council, for instance, brings together 14- to 17-year-olds to serve as advocates for digital well-being and advisers on civility. TikTok and Snap have implemented similar programs, collaborating with organizations like the Digital Wellness Lab at Boston Children's Hospital to create youth advisory boards. These councils provide platforms with valuable first-hand perspectives on how young users interact with their services, helping to inform more effective and relevant safety features and policies.

Studies have shown that including youth in decision-making processes can lead to more responsive policies and stronger partnerships between young people and decision-makers.[164, 165] Additionally, participation in these councils can benefit the youth members themselves, improving their confidence, self-esteem, and sense of purpose, while also developing valuable skills in leadership, public speaking, and policy development. By actively involving young users in the process of creating safer online environments, platforms not only gain essential insights into the lived experiences of young users of their services, but also foster a sense of ownership, proactive involvement, and responsibility among their youth communities. Ideally, this further empowers them to become ambassadors for online safety and security within their peer groups – which can inspire and influence many others beyond the youth council itself.

Generally speaking, UNICEF's 2024 Report on CRIAs in the digital environment revealed that companies are eager to engage with children but face challenges in doing so effectively.[166] Even though some companies utilize youth councils and perform qualitative interviews and focus groups, these methods often lack comprehensive geographic and age-group representation. This hinders the ability of platforms to gather insights relevant to all child users of their services. Resource constraints and time pressures further complicate efforts to conduct meaningful consultations with children across diverse user bases. This can be remedied through consultation and partnerships with external researchers in the field who can create appropriate study samples marked by demographic diversity, geographic representation, and developmental appropriateness.

Cross-Industry Signal Sharing

*(**S**afety and Protection from Harms, **E**vidence-Based Practices, **S**ecurity and Privacy)*

Social media platforms have traditionally operated in silos, in part due to an effort to guard their proprietary information, data, and approaches to content moderation. This protective stance is rooted in the highly competitive nature of the tech industry, where unique algorithms, user engagement strategies, and content policies are seen as key differentiators. Companies may view their Trust and Safety practices as trade secrets, and fear that sharing too much information could give competitors an edge or expose vulnerabilities in their systems that can be exploited or circumvented. While understandable from a business perspective, this approach may hamper the development of industry-wide best practices for addressing online harms.

The reality is that all social media platforms share a common goal: to keep certain types of harmful content off their platforms. Moreover, some collaborative efforts have been in place for years. One of the most prominent tools in the fight against CSEA and CSAM is PhotoDNA, developed by Microsoft and widely adopted across the industry. PhotoDNA creates a unique digital signature (hash) of images, allowing platforms to detect and disrupt the dissemination of known CSAM by comparing uploaded content against a database of previously identified materials. Many companies also participate in initiatives like Project Lantern, which enables the sharing of information about activities and accounts violating policies against online child sexual exploitation and abuse. These existing collaborations have shown the value of cross-industry approaches in tackling such serious issues. Most recently, there is a new development in this space. Thrive, a cross-industry signal-sharing program led by The Mental Health Coalition, was launched in September 2024 and brings together major tech companies like Meta, Snap, and TikTok to collaborate on combating the spread of suicide and self-harm content.[167] (MHC, 2024). By sharing signals about content that violates their respective policies, participating platforms can more effectively identify and address harmful material across various social media ecosystems, thereby reducing the risk of such content slipping through the cracks and harming minors.

The potential of Thrive in mitigating the propagation of suicide and self-harm content can as a template for tackling other forms of online harm. For instance, a signal-sharing program can be built to address issues such as sexually explicit deepfakes, hate speech, school shooting threats, and other clearly problematic content that spreads across multiple platforms. Moreover, this collaborative approach can extend beyond content moderation. Tools for data anonymization, virtual data rooms, and automated governance monitoring can help address concerns about revealing sensitive information or intellectual property while facilitating valuable collaboration. These technologies are making it easier and safer for companies to build trust and share knowledge to tackle thorny problems in youth online safety that they cannot solve alone.

References

1. Bercovici J. Who coined "social media"? Web pioneers compete for credit. Forbes *Disponível*. 2010.
2. Kapoor KK, Tamilmani K, Rana NP, Patil P, Dwivedi YK, Nerur S. Advances in social media research: Past, present and future. *Information Systems Frontiers*. 2018;20:531-558.
3. Aichner T, Grünfelder M, Maurer O, Jegeni D. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking*. 2021;24(4):215-222.
4. Nesi J, Choukas-Bradley S, Prinstein MJ. Transformation of adolescent peer relations in the social media context: Part 1—A theoretical framework and application to dyadic peer relationships. *Clinical Child and Family Psychology Review*. 2018;21:267-294.
5. Pérez-Torres V. Social media: a digital social mirror for identity development during adolescence. *Current Psychology*. 2024:1-11.
6. Raiziene S, Erentaite R, Pakalniskiene V, Grigutyte N, Crocetti E. Identity formation patterns and online activities in adolescence. *Identity*. 2022;22(2):150-165.
7. Wartella E, Rideout V, Montague H, Beaudoin-Ryan L, Lauricella A. Teens, health and technology: A national survey. *Media and Communication*. 2016;4(3):13-23.
8. Hinduja S, Patchin JW. Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. *Journal of Adolescence.* 2008;31(1):125-146.
9. Uhls YT, Ellison NB, Subrahmanyam K. Benefits and costs of social media in adolescence. *Pediatrics*. 2017;140(Supplement_2):S67-S70.
10. boyd d. Identity Production in a Networked Culture: Why Youth Heart MySpace. St. Louis, Missouri; American Association for the Advancement of Science Annual Conference, 2006.
11. boyd d. *It's Complicated: The Social Lives of Networked Teens*. Yale University Press; 2014.
12. Craig SL, McInroy L, McCready LT, Alaggia R. Media: A catalyst for resilience in lesbian, gay, bisexual, transgender, and queer youth. *Journal of LGBT Youth*. 2015;12(3):254-275.
13. Tynes BM, Umana-Taylor AJ, Rose CA, Lin J, Anderson CJ. Online racial discrimination and the protective function of ethnic identity and self-esteem for African American adolescents. *Developmental Psychology*. 2012;48(2):343.
14. Raghavendra P, Newman L, Grace E, Wood D. 'I could never do that before': effectiveness of a tailored Internet support intervention to increase the social participation of youth with disabilities. *Child: Care, Health and Development*. 2013;39(4):552-561.
15. Haimson OL, Brubaker JR, Dombrowski L, Hayes GR. Disclosure, stress, and support during gender transition on Facebook. 2015:1176-1190.
16. Macintosh L, Bryson M. Youth, MySpace, and the interstitial spaces of becoming and belonging. *Journal of LGBT Youth*. 2008;5(1):133-142.
17. Sapiro B, Ward A. Marginalized youth, mental health, and connection with others: A review of the literature. *Child and Adolescent Social Work Journal*. 2020;37(4):343-357.
18. Charmaraman L, Hernandez JM, Hodes R. Marginalized and understudied populations using digital media. *Handbook of Adolescent Digital Media Use and Mental Health*. 2022:188-214.
19. Jenkins H, Shresthova S, Gamber-Thompson L, Kligler-Vilenchik N, Zimmerman A. *By any media necessary: The New Youth Activism*. New York University Press; 2016.
20. Wright K, McLeod J. Activism, Rights and Hope: Young People and Their Advocates Mobilising for Social Change. *Childhood, Youth and Activism: Demands for Rights and Justice from Young People and their Advocates*. Emerald Publishing Limited; 2023:1-18.
21. Mihailidis P. The civic potential of memes and hashtags in the lives of young people. *Discourse: Studies in the Cultural Politics of Education*. 2020;41(5):762-781.
22. Neag A, Supa M, Mihailidis P. Researching Social Media and Activism With Children and Youth: A Scoping Review. *International Journal of Communication*. 2024;18:22.
23. Mundt M, Ross K, Burnett CM. Scaling social movements through social media: The case of Black Lives Matter. *Social Media + Society*. 2018;4(4):2056305118807911.
24. Haugestad CA, Skauge AD, Kunst JR, Power SA. Why do youth participate in climate activism? A mixed-methods investigation of the# FridaysForFuture climate protests. *Journal of Environmental Psychology*. 2021;76:101647.
25. O'Brien K, Selboe E, Hayward BM. Exploring youth activism on climate change. *Ecology and Society*. 2018;23(3).
26. Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries. 2011.
27. Livingstone S, Stoilova M. The 4Cs: Classifying online risk to children. *CO:RE Short Report Series on Key Topics*. Leibniz-Institut für Medienforschung, Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence; 2021. https://doi.org/10.21241/ssoar.71817.
28. Marino C, Canale N, Melodia F, Spada MM, Vieno A. The overlap between problematic smartphone use and problematic social media use: a systematic review. *Current Addiction Reports*. 2021:1-12.
29. Sherman LE, Hernandez LM, Greenfield PM, Dapretto M. What the brain 'Likes': neural correlates of providing feedback on social media. *Social Cognitive and Affective Neuroscience*. 2018;13(7):699-707.
30. Rosenthal-von der Pütten AM, Hastall MR, Köcher S, et al. "Likes" as social rewards: Their role in online social comparison and decisions to like other People's selfies. *Computers in Human Behavior*. 2019;92:76-86.
31. Mujica AL, Crowell CR, Villano MA, Uddin KM. Addiction by design: Some dimensions and challenges of excessive social media use. *Medical Research Archives*. 2022;10(2).
32. Montag C, Lachmann B, Herrlich M, Zweig K. Addictive features of social media/messenger platforms and freemium games against the background of psychological and economic theories. *International Journal of Environmental Research and Public Health*. 2019;16(14):2612.

33. Lee U, Lee J, Ko M, et al. Hooked on smartphones: an exploratory study on smartphone overuse among college students. 2014:2327-2336.
34. Koepp MJ, Gunn RN, Lawrence AD, et al. Evidence for striatal dopamine release during a video game. *Nature*. 1998;393(6682):266-268.
35. Zald DH, Boileau I, El-Dearedy W, et al. Dopamine transmission in the human striatum during monetary reward tasks. *Journal of Neuroscience*. 2004;24(17):4105-4112.
36. Sherman LE, Payton AA, Hernandez LM, Greenfield PM, Dapretto M. The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Sciences*. 2016;27(7):1027-1035.
37. Pellegrino A, Stasi A, Bhatiasevi V. Research trends in social media addiction and problematic social media use: A bibliometric analysis. *Frontiers in Psychiatry*. 2022;13:1017506.
38. Swart J. Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media + Society*. 2021;7(2):20563051211008828.
39. Rubenking B, Bracken CC, Sandoval J, Rister A. Defining new viewing behaviours: What makes and motivates TV binge-watching? *International Journal of Digital Television*. 2018;9(1):69-85.
40. Alter A. *Irresistible: The Rise of Addictive Technology and The Business of Keeping Us Hooked*. Penguin; 2017.
41. Georges PM, Bayle-Tourtoulou A-S, Badoc M. *Neuromarketing in action: How to Talk and Sell to the Brain*. Kogan Page Publishers; 2013.
42. Alexander J. Netflix is letting people watch things faster or slower with new playback speed controls. *The Verge*. 2020;31
43. Kumpf B, Hanson A. *Reshaping Social Media: from Persuasive Technology to Collective Intelligence*. 2021.
44. Anderson K, Burford O, Emmerton L. Mobile health apps to facilitate self-care: a qualitative study of user experiences. *PloS One*. 2016;11(5):e0156164.
45. Villalobos-Zúñiga G, Cherubini M. Apps that motivate: a taxonomy of app features based on self-determination theory. *International Journal of Human-Computer Studies*. 2020;140:102449.
46. Baggio S, Starcevic V, Studer J, et al. Technology-mediated addictive behaviors constitute a spectrum of related yet distinct conditions: A network perspective. *Psychology of Addictive Behaviors*. 2018;32(5):564.
47. Baggio S, Starcevic V, Billieux J, et al. Testing the spectrum hypothesis of problematic online behaviors: A network analysis approach. *Addictive Behaviors*. 2022;135:107451.
48. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*. 2018;6(1):3-17.
49. Fink E, Patalay P, Sharpe H, Holley S, Deighton J, Wolpert M. Mental health difficulties in early adolescence: a comparison of two cross-sectional studies in England from 2009 to 2014. *Journal of Adolescent Health*. 2015;56(5):502-507.
50. Kim Y, Hagquist C. Trends in adolescent mental health during economic upturns and downturns: a multilevel analysis of Swedish data 1988-2008. *Journal of Epidemiology and Community Health*. 2018;72(2):101-108.
51. Wiens K, Williams JV, Lavorato DH, et al. Is the prevalence of major depression increasing in the Canadian adolescent population? Assessing trends from 2000 to 2014. *Journal of Affective Disorders*. 2017;210:22-26.
52. Duinhof EL, Stevens GW, Van Dorsselaer S, Monshouwer K, Vollebergh WA. Ten-year trends in adolescents' self-reported emotional and behavioral problems in the Netherlands. *European Child & Adolescent Psychiatry*. 2015;24:1119-1128.
53. von Soest T, Wichstrøm L. Secular trends in depressive symptoms among Norwegian adolescents from 1992 to 2010. *Journal of Abnormal Child Psychology*. 2014;42:403-415.
54. Potrebny T, Wiium N, Lundegård MM-I. Temporal trends in adolescents' self-reported psychosomatic health complaints from 1980-2016: A systematic review and meta-analysis. *PloS One*. 2017;12(11):e0188374.
55. Cosma A, Stevens G, Martin G, et al. Cross-national time trends in adolescent mental well-being from 2002 to 2018 and the explanatory role of schoolwork pressure. *Journal of Adolescent Health*. 2020;66(6):S50-S58.
56. Bor W, Dean AJ, Najman J, Hayatbakhsh R. Are child and adolescent mental health problems increasing in the 21st century? A systematic review. *Australian & New Zealand Journal of Psychiatry*. 2014;48(7):606-616.
57. Collishaw S. Annual research review: secular trends in child and adolescent mental health. *Journal of Child Psychology and Psychiatry*. 2015;56(3):370-393.
58. Keyes KM, Gary D, O'Malley PM, Hamilton A, Schulenberg J. Recent increases in depressive symptoms among US adolescents: trends from 1991 to 2018. *Social Psychiatry and Psychiatric Epidemiology*. 2019;54:987-996.
59. Vuorre M, Przybylski AK. Global well-being and mental health in the internet age. *Clinical Psychological Science*. 2023:21677026231207791.
60. Orben A, Przybylski AK, Blakemore S-J, Kievit RA. Windows of developmental sensitivity to social media. *Nature Communications*. 2022;13(1):1649.
61. Benton TD, Boyd RC, Njoroge WF. Addressing the global crisis of child and adolescent mental health. *JAMA Pediatrics.* 2021;175(11):1108-1110.
62. von Soest T, Kozák M, Rodríguez-Cano R, et al. Adolescents' psychosocial well-being one year after the outbreak of the COVID-19 pandemic in Norway. *Nature Human Behaviour*. 2022;6(2):217-228.
63. de Abreu PME, Neumann S, Wealer C, Abreu N, Macedo EC, Kirsch C. Subjective well-being of adolescents in Luxembourg, Germany, and Brazil during the COVID-19 pandemic. *Journal of Adolescent Health*. 2021;69(2):211-218.
64. Low N, Mounts NS. Economic stress, parenting, and adolescents' adjustment during the COVID-19 pandemic. *Family Relations*. 2022;71(1):90-107.
65. Munasinghe S, Sperandei S, Freebairn L, et al. The impact of physical distancing policies during the COVID-19 pandemic on health and well-being among Australian adolescents. *Journal of Adolescent Health*. 2020;67(5):653-661.
66. Wiguna T, Anindyajati G, Kaligis F, et al. Brief research report on adolescent mental well-being and school closures during the COVID-19 pandemic in Indonesia. *Frontiers in Psychiatry*. 2020;11:598756.
67. Odd D, Williams T, Appleby L, Gunnell D, Luyt K. Child suicide rates during the COVID-19 pandemic in England. *Journal of Affective Disorders Reports*. 2021;6:100273.
68. Levita L, Miller JG, Hartman TK, et al. Report1: Impact of Covid-19 on young people aged 13-24 in the UK-preliminary findings. 2020.

69. Ellis WE, Dumas TM, Forbes LM. Physically isolated but socially connected: Psychological adjustment and stress among adolescents during the initial COVID-19 crisis. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*. 2020;52(3):177.

70. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*. 2015;56(3):345-365.

71. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. *JAMA Pediatrics*. 2021;175(11):1142-1150.

72. Talamonti D, Schneider J, Gibson B, Forshaw M. The impact of national and international financial crises on mental health and well-being: a systematic review. *Journal of Mental Health*. 2023:1-38.

73. Prime H, Wade M, Browne DT. Risk and resilience in family well-being during the COVID-19 pandemic. *American Psychologist*. 2020;75(5):631.

74. Bell DN, Blanchflower DG. US and UK labour markets before and during the Covid-19 crash. *National Institute Economic Review*. 2020;252:R52-R69.

75. Yoo N, Jang SH. Perceived household financial decline and physical/mental health among adolescents during the COVID-19 crisis: Focusing on gender differences. *Preventive Medicine Reports*. 2023;32:102119.

76. Park CL, Russell BS, Fendrich M, Finkelstein-Fox L, Hutchison M, Becker J. Americans' COVID-19 stress, coping, and adherence to CDC guidelines. *Journal of General Internal Medicine*. 2020;35:2296-2303.

77. Cluver L, Lachman JM, Sherr L, et al. Parenting in a time of COVID-19. 2020.

78. Russell BS, Hutchison M, Tambling R, Tomkunas AJ, Horton AL. Initial challenges of caregiving during COVID-19: Caregiver burden, mental health, and the parent–child relationship. *Child Psychiatry & Human Development*. 2020;51(5):671-682.

79. Lancet T. An age of uncertainty: mental health in young people. 2022;400(10352):539.

80. Schweizer S, Lawson RP, Blakemore S-J. Uncertainty as a driver of the youth mental health crisis. *Current Opinion in Psychology*. 2023;53:101657.

81. Sugg MM, Wertis L, Ryan SC, Green S, Singh D, Runkle JD. Cascading disasters and mental health: The February 2021 winter storm and power crisis in Texas, USA. *Science of the Total Environment*. 2023;880:163231.

82. Clark LS. *The parent app: Understanding Families in the Digital Age*. Oxford University Press; 2012.

83. Vickery JR. *Worried about the wrong things: Youth, Risk, and Opportunity in the Digital World*. MIT Press; 2017.

84. Livingstone S, Blum-Ross A. *Parenting for a Digital Future: How Hopes and Fears About Technology Shape Children's Lives*. Oxford University Press, USA; 2020.

85. Smahel D, Machackova H, Mascheroni G, et al. EU Kids Online 2020: Survey results from 19 countries. EU Kids Online; 2020. Accessed November 15, 2023. http://hdl.handle.net/20.500.12162/5299.

86. Livingstone S, Stoilova M. The 4Cs: Classifying online risk to children. 2021.

87. Stoilova M, Bulger M, Livingstone S. Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use. *Journal of Children and Media*. 2024;18(1):29-49.

88. Livingstone S. Online risk, harm and vulnerability: Reflections on the evidence base for child Internet safety policy. *ZER: Journal of Communication Studies*. 2013;18(35):13-28.

89. Livingstone S, Ólafsson K, Helsper EJ, Lupiáñez-Villanueva F, Veltri GA, Folkvord F. Maximizing opportunities and minimizing risks for children online: The role of digital skills in emerging strategies of parental mediation. *Journal of Communication*. 2017;67(1):82-105.

90. Nichols S, Selim N. Digitally mediated parenting: A review of the literature. *Societies*. 2022;12(2):60.

91. Elsaesser C, Russell B, Ohannessian CM, Patton D. Parenting in a digital age: A review of parents' role in preventing adolescent cyberbullying. *Aggression and Violent Behavior*. 2017;35:62-72.

92. Busso D, Cakmakli A, Munger J. How we've used co-design to develop parental supervision tools at Meta. Meta; 2022.

93. Maria N. The new European strategy for a better internet for kids (BIK+). EPRS: European Parliamentary Research Service; 2022. September 28, 2022. https://coilink.org/20.500.12592/hfnrkb.

94. DTSP. Age Assurance: Guiding Principles and Best Practices. Digital Trust & Safety Partnership; 2023. Accessed September 30, 2023. https://dtspartnership.org/wp-content/uploads/2023/09/DTSP_Age-Assurance-Best-Practices.pdf.

95. Ofcom. Protecting children from harms online. *Volume 3: The Causes and Impacts of Online Harms to Children*. United Kingdom Office of Communications; 2024. May 8, 2024. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol3-causes-impacts-of-harms-to-children.pdf?v=336052.

96. 5Rights Foundation. A High Level of Privacy, Safety & Security for Minors – A best practices baseline for the implementation of the Digital Services Act for Children. Eurochild and European Parliament Intergroup on Child Rights; 2024. February 2024. https://5rightsfoundation.com/wp-content/uploads/2024/08/5rights-foundation-a-high-level-of-privacy-safety-and-security-for-minors-dsa-baseline-2024-final-1.pdf.

97. ICO. Age Appropriate Design: A Code of Practicefor Online Services. Information Commissioner's Office; 2020. September 2, 2020. https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services-2-1.pdf.

98. Eremin AA. Why Prohibition Tactics do not Work? A Critical Evaluation of Historic Experience of the United States in Fighting Alcohol and Drugs. 2022.

99. Hicks DC, Petrunik MG. The Best Intentions Are Not Enough: Drug Prohibition as a Failed Intervention Strategy. *Canadian Review of Social Policy*. 1997:1-17.

100. Vanwesenbeeck I. Sex work criminalization is barking up the wrong tree. *Archives of Sexual Behavior*. 2017;46(6):1631-1640.

101. Mueller ML. *Networks and States: The Global Politics of Internet Governance*. MIT Press; 2010.

102. Kleck G. *Point Blank: Guns and violence in America*. Routledge; 2017.

103. Werb D, Mills EJ, DeBeck K, Kerr T, Montaner JS, Wood E. The effectiveness of anti-illicit-drug public-service announcements: a systematic review and meta-analysis. *Journal of Epidemiology and Community Health*. 2011;65(10):834-840.

104. Platt L, Grenfell P, Meiksin R, et al. Associations between sex work laws and sex workers' health: A systematic review and meta-analysis of quantitative and qualitative studies. *PloS Medicine*. 2018;15(12):e1002680.

105. Zittrain J. *The Future of the Internet: and How to Stop It*. Penguin UK; 2009.

106. Király O, Griffiths MD, King DL, et al. Policy responses to problematic video game use: A systematic review of current measures and future possibilities. *Journal of Behavioral Addictions*. 2018;7(3):503-517.

107. Miller AC. # DictatorErdogan: How social media bans trigger backlash. *Political Communication*. 2022;39(6):801-825.

108. Huddleston J. Would New Legislation Actually Make Kids Safer Online? Analyzing the Consequences of Recent Youth Online Safety Proposals. *Cato Institute Briefing Paper*. 2023;(150).

109. APA. APA resolution on violent video games. February 2020 Revision to the 2015 Resolution. American Psychological Association; 2020. https://www.apa.org/about/policy/resolution-violent-video-games.pdf.

110. Mills K. APA reaffirms position on violent video games and violent behavior. American Psychological Association; 2020. March 3, 2020. https://www.apa.org/news/press/releases/2020/03/violent-video-games-behavior.

111. Australian Government Attorney General’s Department. Literature review on the impact of playing violent video games on aggression. Commonwealth of Australia Barton, Australian Capital Territory, Australia; 2010.

112. SMC. Summary of violent computer games and aggression - an overview of the research 2000-2011. Swedish Media Council,; 2012. https://videogameseurope.eu/wp-content/uploads/2012/01/literature_review_violent_games_-_summary.pdf.

113. Przybylski AK, Weinstein N. Violent video game engagement is not associated with adolescents’ aggressive behaviour: evidence from a registered report. *Royal Society Open Science*. 2019;6(2):171474.

114. Ybarra ML, Diener-West M, Markow D, Leaf PJ, Hamburger M, Boxer P. Linkages between internet and other media violence with seriously violent behavior by youth. *Pediatrics*. 2008;122(5):929-937.

115. Surette R, Maze A. Video game play and copycat crime: An exploratory analysis of an inmate population. *Psychology of Popular Media Culture*. 2015;4(4):360.

116. DeCamp W. Impersonal agencies of communication: Comparing the effects of video games and other risk factors on violence. *Psychology of Popular Media Culture*. 2015;4(4):296.

117. Ferguson CJ, Klisinan D, Hogg JL, et al. Societal Violence and Video Games: Public Statements of a Link are Problematic. *The Amplifier Magazine*. Society for Media Psychology and Technology, Division 46 of the American Psychological Association; 2017. https://div46amplifier.com/2017/06/12/news-media-public-education-and-public-policy-committee/.

118. Király O, Nagygyörgy K, Griffiths MD, Demetrovics Z. Problematic online gaming. *Behavioral Addictions*. Elsevier; 2014:61-97.

119. Männikkö N, Billieux J, Kääriäinen M. Problematic digital gaming behavior and its relation to the psychological, social and physical health of Finnish adolescents and young adults. *Journal of Behavioral Addictions*. 2015;4(4):281-288.

120. Demetrovics Z, Urbán R, Nagygyörgy K, et al. The development of the problematic online gaming questionnaire (POGQ). *PloS One*. 2012;7(5):e36417.

121. Pontes HM, Macur M, Griffiths MD. Internet gaming disorder among Slovenian primary schoolchildren: Findings from a nationally representative sample of adolescents. *Journal of Behavioral Addictions*. 2016;5(2):304-310.

122. Humphreys K, McLellan AT. A policy-oriented review of strategies for improving the outcomes of services for substance use disorder patients. *Addiction*. 2011;106(12):2058-2066.

123. Choi J, Cho H, Lee S, Kim J, Park E-C. Effect of the online game shutdown policy on internet use, internet addiction, and sleeping hours in Korean adolescents. *Journal of Adolescent Health*. 2018;62(5):548-555.

124. Zendle D, Flick C, Gordon-Petrovskaya E, Ballou N, Xiao LY, Drachen A. No evidence that Chinese playtime mandates reduced heavy gaming in one segment of the video games industry. *Nature Human Behaviour*. 2023;7(10):1753-1766.

125. Davies B, Blake E. Evaluating existing strategies to limit video game playing time. I*EEE Computer Graphics and Applications*. 2016;36(2):47-57.

126. O’Neill B, Dopona V. The Better Internet for Kids Policy Monitor Report. European Schoolnet; 2024. May 11, 2024. https://better-internet-for-kids.europa.eu/sites/default/files/documents/167024/7159869/BIK%20Policy%20Monitor%20Report%202024.pdf.

127. Ofcom. A feasibility study of using Well-being metrics to evaluate outcomes in Online Safety. *Economics Discussion Paper Series Issue 13*. United Kingdom Office of Communications; 2024. October 8, 2024. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/evaluating-wellbeing-impacts-edp/a-feasibility-study-of-using-wellbeing-metrics-to-evaluate-outcomes-in-online-safety.pdf?v=382566.

128. Collier A, Noula I. Five Cautionary Notes For Successful Implementation of the DSA for Children’s Best Interests. Tech Policy Press; 2024. October 31, 2024. https://www.techpolicy.press/five-cautionary-notes-for-successful-implementation-of-the-dsa-for-childrens-best-interests/.

129. Ofcom. Children and parents: media use and attitudes report 2024. *Making Sense of Media*. United Kingdom Office of Communications; 2024. April 19, 2024. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/children/children-media-use-and-attitudes-2024/childrens-media-literacy-report-2024.pdf?v=368229.

130. Kobilke L, Markiewitz A. The Momo Challenge: measuring the extent to which YouTube portrays harmful and helpful depictions of a suicide game. *SN Social Sciences*. 2021;1:1-30.

131. Lorenz T. How a toilet-themed YouTube series became the biggest thing online. Washington Post; 2023. December 10, 2023. https://www.washingtonpost.com/technology/2023/12/10/skibidi-toilets-you-tube-children-internet/.

132. Marwick AE. To catch a predator? The MySpace moral panic. *First Monday*. 2008.

133. Quayle E. Internet risk research and child sexual abuse: a misdirected moral panic? *Revisiting Moral Panics*. Policy Press; 2015:103-112.

134. Anda F, Dixon E, Bou-Harb E, Le-Khac N-A, Scanlon M. Vec2UAge: Enhancing underage age estimation performance through facial embeddings. *Forensic Science International: Digital Investigation*. 2021;36:301119.

135. Jarvie C, Renaud K. Are you over 18? A snapshot of current age verification mechanisms. 2021.

136. CSI. Protecting Children Online: Mandatory Device-Based Age Verification and Parental Controls. Crime Stoppers International; 2024. December, 2024. https://static1.squarespace.com/static/6354fb8a954945102e2819e0/t/67650016ca459b0b0b955185/1734672499996/POSITION+PAPER+Device-Based+Age+Verification.pdf.

137. ICMEC. Statement on Age Verification. International Centre for Missing & Exploited Children; 2024. June 27, 2024. https://www.icmec.org/press/statement-on-age-verification/.

138. Michael K. Mitigating Risk and Ensuring Human Flourishing Using Design Standards: IEEE 2089-2021 an Age Appropriate Digital Services Framework for Children. *IEEE Transactions on Technology and Society*. 2024.

139. Primack BA, Bisbey MA, Shensa A, et al. The association between valence of social media experiences and depressive symptoms. *Depression and Anxiety*. 2018;35(8):784-794.

140. Ebel E, Kellner M, Möller T, Rupalla F, Zuyeva D. Hands off: Consumer perceptions of advanced driver assistance systems. McKinsey & Company; 2023. July 19, 2023. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/hands-off-consumer-perceptions-of-advanced-driver-assistance-systems.

141. WEF. Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms. World Economic Forum; 2023. August 4, 2023. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

142. Patchin JW, Hinduja S. Tween Cyberbullying in 2020. Cartoon Network, Warner Media.; 2020.

143. Patchin JW, Hinduja S. Words Wound: *Delete Cyberbullying and Make Kindness Go Viral*. Minneapolis, MN; 2014.

144. Bucher T, Helmond A. The affordances of social media platforms. *The SAGE Handbook of Social Media*. 2018;1:233-254.

145. Hinduja S, Patchin JW. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*. 3rd ed. Sage Publications; 2024.

146. Montada L. Injustice in harm and loss. *Social Justice Research*. 1994;7(1):5-28.

147. Wemmers J-A. Victims' experiences in the criminal justice system and their recovery from crime. *International Review of Victimology*. 2013;19(3):221-233.

148. Campbell R, Raja S. Secondary victimization of rape victims: Insights from mental health professionals who treat survivors of violence. *Violence and Victims*. 1999;14(3):261-275.

149. Patterson D. The linkage between secondary victimization by law enforcement and rape case outcomes. *Journal of Interpersonal Violence*. 2011;26(2):328-347.

150. Wemmers J-A. Restorative justice for victims of crime: A victim-oriented approach to restorative justice. *International Review of Victimology*. 2002;9(1):43-59.

151. Garvin M, Beloof DE. Crime victim agency: Independent lawyers for sexual assault victims. *Ohio State Journal of Criminal Law*. 2015;13:67.

152. Garvin M, LeClaire S. Polyvictims: Victims' rights enforcement as a tool to mitigate "secondary victimization" in the criminal justice system. *National Crime Victim Law Institute Victim Law Bulletin*. 2013.

153. Orth U. Secondary victimization of crime victims by criminal proceedings. *Social Justice Research*. 2002;15(4):313-325.

154. Jhaver S, Chen QZ, Knauss D, Zhang AX. Designing word filter tools for creator-led comment moderation. 2022:1-21.

155. Paul K. Meta overhauls rules on deepfakes, other altered media. Reuters; 2024. April 5, 2024. Accessed April 5, 2024. https://www.reuters.com/technology/cybersecurity/meta-overhauls-rules-deepfakes-other-altered-media-2024-04-05/.

156. Wardle C, Derakhshan H. Information disorder: *Toward an Interdisciplinary Framework for Research and Policymaking*. vol 27. Council of Europe Strasbourg; 2017.

157. Gillespie T. Custodians of the Internet: *Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press; 2018.

158. van de Weijer S, Leukfeldt R, Van der Zee S. Reporting cybercrime victimization: determinants, motives, and previous experiences. *Policing: An International Journal*. 2020;43(1):17-34.

159. Pezzella FS, Fetzer MD, Keller T. The dark figure of hate crime underreporting. *American Behavioral Scientist*. 2019:0002764218823844.

160. Kidd RF, Chayet EF. Why do victims fail to report? The psychology of criminal victimization. *Journal of Social Issues*. 1984;40(1):39-50.

161. Hinduja S, Patchin JW. Cyberbullying Through the Lens of Trauma: An Empirical Examination of US Youth. *[redacted]*. In review.

162. Project Rockit. Shaping our Feeds: Young People's Experience of Social Media Algorithms. 2024. November 7, 2024. https://bit.ly/4hsvbjg.

163. Celadin T, Panizza F, Capraro V. Promoting civil discourse on social media using nudges: A tournament of seven interventions. *PNAS Nexus*. 2024;3(10)doi:https://doi.org/10.1093/pnasnexus/pgae380.

164. OECD. Engaging youth in policy-making processes (Module 6). Organisation for Economic Cooperation and Development; 2017. https://doi.org/10.1787/9789264283923-10-en.

165. OECD. Evidence-based Policy Making for Youth Well-being-A Toolkit. Organisation for Economic Cooperation and Development; 2017. https://doi.org/10.1787/9789264283923-en.

166. UNICEF. Child Rights Impact Assessments in Relation to the Digital Environment. UNICEF; 2024. https://www.unicef.org/media/156046/file/Child%20Rights%20Impact%20Assessments%20in%20Relation%20to%20the%20Digital%20Environment.pdf.

167. MHC. The First Suicide and Self-Harm Cross-Industry Signal Sharing Program to be Established Under the Leadership of The Mental Health Coalition. Mental Health Coalition; 2024. https://www.prnewswire.com/news-releases/the-first-suicide-and-self-harm-cross-industry-signal-sharing-program-to-be-established-under-the-leadership-of-the-mental-health-coalition-302246042.html.