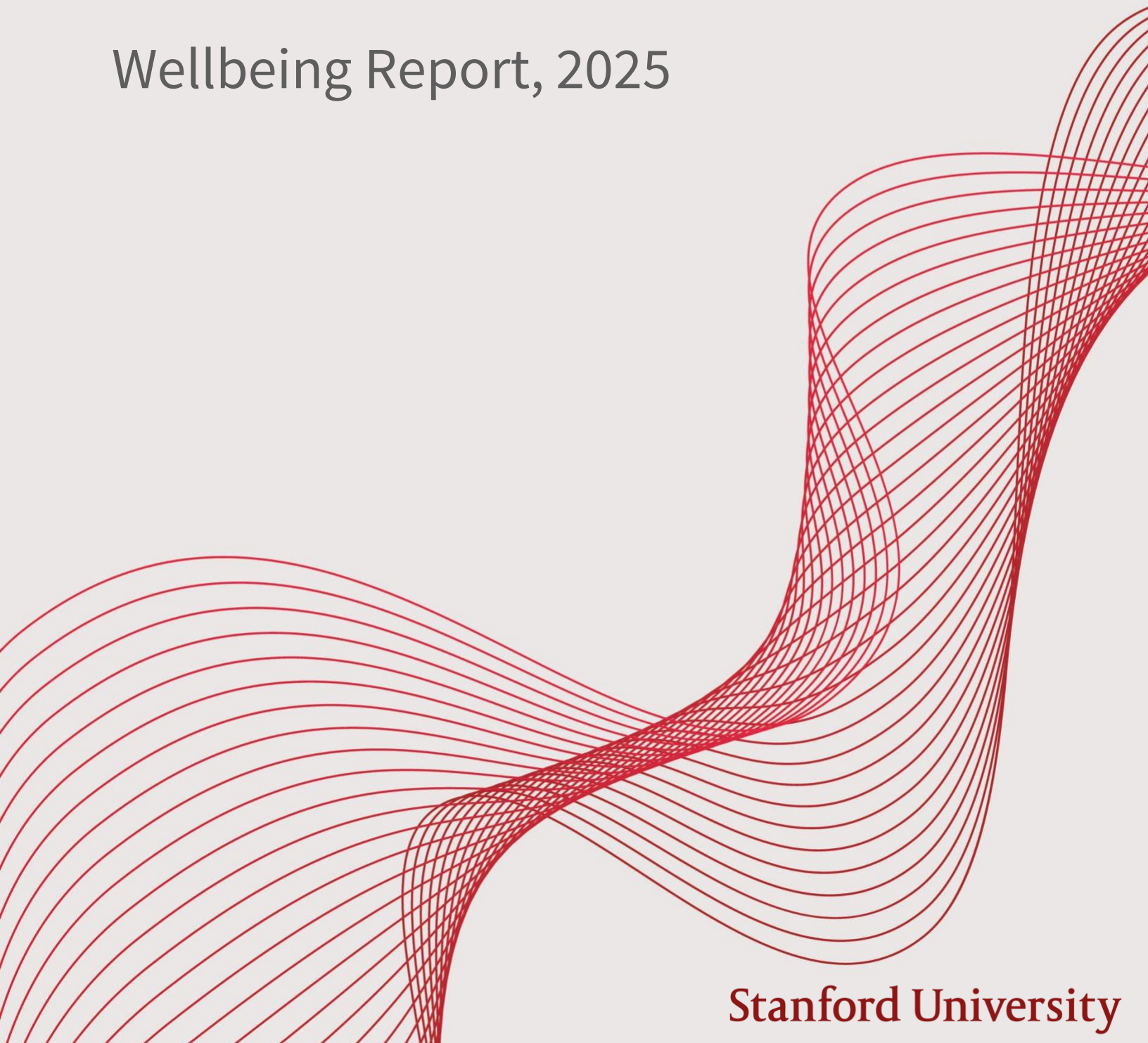
Stanford Youth Safety and Digital Wellbeing Report, 2025



Stanford University

Introduction

The conversation around social media and youth wellbeing has grown increasingly complex, as researchers, policymakers, and industry leaders work to understand the true scope of potential harms and benefits. While new regulatory frameworks, such as the EU Digital Services Act (DSA), mandate risk-based approaches to assessing and mitigating online harms, many questions remain about how these risks are defined, measured, and addressed in practice. How do different harms impact young people? What strategies are most effective in reducing risks while preserving the positive aspects of social media use? And how can we balance safety measures with broader considerations including privacy, free speech, and technological innovation?

To explore these questions, the Center for Digital Health (CDH) and Social Media Lab (SML) at Stanford University convened a workshop bringing together experts from diverse fields, including public health, education, mental health, and industry. The objectives of the workshop were: 1) to establish a prioritized list of youth-specific social media harms that have been considered in the scientific literature; and 2) for each possible harm, to explore practical strategies for measuring the scale of the issue and to document evidence-based mitigation approaches.

This report summarizes the key insights from the workshop, shedding light on the evolving landscape of youth digital wellbeing. Rather than assuming a singular narrative or that the evidence is conclusive regarding a specific harm, the discussion sought to critically examine the varying degrees of risk associated with different possible youth-specific harms, the trade-offs of different mitigation approaches, and opportunities for both regulatory and voluntary solutions. Its findings will enable better understanding not only the challenges but also the potential for shaping a digital environment that supports the safety and wellbeing of youth and adolescent users. Throughout this report, the use of the term 'harm(s)' refers to the *potential harms* associated with social media, as rasied and examined in the scientific literature.

Taxonomy of Harms

In order to develop effective safeguards for youth users of social media platforms, it is essential that we move beyond only debating the aggregate impact of social media on health outcomes. Concerns regarding social media harms are diverse, with distinct causes, effects, and implications. Aggregated measures—while important to inform overall priority setting—are not designed to capture these nuances; only focusing on these measures may cause us to overlook or misestimate individual risks and ways to address them. Each possible harm should be considered individually, with tailored measurement and mitigation strategies. The Integrated Harm Framework (IHF) outlined below provides a foundation for this process.

A preliminary list of potential harms identified in the scientific literature was collated from four sources: the World Economic Forum's 'Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms' (2023); the National Academies' 'Social Media and Adolescent Health' report (2024); the Stanford Social Media Lab's 'Adolescents and Well-being Report' (2024); and the Minnesota Attorney General's 'Report on Emerging Technology and Its Effects on Youth Well-Being' (2024). These four sources were selected to provide coverage of possible social media harms from different perspectives. The World Economic Forum's report presents a typology that distinguishes risks driven by *content* (content production, distribution, and consumption), *contact* (online interactions with others), *conduct* (behaviour facilitated by technology), and *contract* (commercialization and datafication). The National Academies' report focuses on the scientific evidence of online harassment, including cyberbullying and sexual offenses. Stanford Social Media Lab's report listed the kinds of harms parents should be aware of from parents' perspectives. The Minnesota Attorney General's described several harms that may be facilitated by technological design choices of technology platforms. For the workshop, these harms were classified into the following four conceptual categories:

1. **Threats to safety/criminal activity**
2. **Health & wellbeing**
3. **Other content-driven harms**
4. **Other harms**

Prior to the workshop, attending experts were surveyed to achieve two objectives: first, to encourage participants to review and reflect on the harms that would be discussed during the workshop; and second, to gather their perspectives on the existing harm categories, as well as recommendations for any additional harms not currently covered in the typology. Participants rated on a 9-point scale how strongly they would recommend including each harm in risk assessments of youth wellbeing and social media (1 = strongly against inclusion, 9 = strongly for inclusion). Survey respondents could also suggest the addition of further possible harms not covered by the preliminary list. Workshop discussion targeted those harms with a mean of less than 8 in this preliminary vote, and a second round of voting was held following the discussions.

The Integrated Harm Framework (IHF) emerged as a unifying model, offering a shared foundation for developing evidence-based and actionable strategies. The IHF including the full list of specific possible harms is presented in **Table 1**, along with mean rating scores and their standard deviations. In total, 22 harms were considered, with mean rating scores ranging from 5.27 to 8.91. The top five harms with the highest mean values were: (1) Adult-minor solicitation, image-based sexual abuse, sextortion; (2) Child Sexual Abuse Material (CSAM) and Child Sexual Exploitation Material (CSEM); (3) Communities and/or content that promote self-harm or suicide; bullying, harassment and stalking (including technology facilitated abuse and gender-based violence); and (4) Communities and/or content that promote eating disorders, dysmorphia, unhealthy body image. The five harms with the lowest mean values were: (1) Parent use of social media and related stress/displacement of social interactions; (2) Algorithmic biases and risks (including recommendations of problematic content or discrimination in decision-making); (3) Infringement of Child Rights: over-limiting child access to information; misinformation and disinformation; and (4) Psychological impacts, including depression, sadness, anxiety, loneliness, lower positive well-being indicators, such as happiness, self-esteem.

For 17 of the 22 harms, the mode was 9. The five harms that did not have a mode of 9 were: Fraud (including identity theft, impersonation, scams); psychological impacts, including depression, sadness, anxiety, loneliness, lower positive well-being indicators, such as happiness, self-esteem; misinformation and disinformation; infringement of Child Rights: over-limiting child access to information; algorithmic biases and risks (including recommendations of problematic content or discrimination in decision-making).

For all harms, a majority of votes were cast in favour of inclusion (i.e. a majority of participants voted for a score of ≥ 5), reflecting that all harms were considered important. It is worth noting that some participants' votes were influenced by their considerations of the prevalence and severity of the harms. Lower mean scores may reflect a participant's perception of how prevalent or severe a harm is, rather than questioning its importance for consideration. Key points of discussion specific to each harm are also included in the table—it is important to note these were not necessarily consensus points.

.

Conceptual Categories:

1. Threats to safety and/or criminal activity	2. Health and wellbeing	3. Other content-driven harms	4. Other harms

	Specific Harm	Mean	SD	Mode	Key points of discussion
1	Adult-minor solicitation, image-based sexual abuse, sextortion	8.91	0.28	9	- Solicitation is among the most severe harms. - While related to **2**, unlike CSAM the content alone is not illegal and necessitates distinct mitigation strategies, making separation from **2** advisable.
2	Child Sexual Abuse Material (CSAM) and Child Sexual Exploitation Material (CSEM)	8.87	0.45	9	- While not highly prevalent, this is among the most severe harms. - Suggestions to consider self-generated CSAM as a separate category.
3	Communities/content that promote self-harm or suicide	8.78	0.72	9	- High severity, though prevalence is low. - Rarity makes measurement difficult. - Recommendation to combine with **5**, **15**.
4	Bullying, harassment and stalking (including technology facilitated abuse and gender-based violence)	8.7	0.62	9	- Affects approximately 1 in 6 youth, although severity varies. - Distinctions between peer-to-peer bullying (which can originate in school) and other forms are important.
5	Communities/content that promote eating disorders, dysmorphia, unhealthy body image	8.52	0.88	9	- Highest-severity forms of this are relatively low prevalence. - Some harmful material can be disguised as wellness content. - Recommendation to combine with **5**, **15**.
6	Exposure to unwanted violent or graphic imagery	8.52	0.88	9	- Suggested as an additional category, or in combination with **9** under a broader category of "unwanted content".
7	Communities/content that promote terrorism or violence	8.35	0.91	9	- Concerns were raised about terminology. Replacing "promoting" with "facilitating recruitment for" was suggested.
8	Privacy (including unintended or unwanted exposure of personal information)	8.33	0.94	9	- User perspectives highlight privacy as a top priority. - Harm can stem from user misunderstandings of platforms' different settings. - Suggestion to separate concerns about i) platform data collection/use, and ii) exposure of information to other users.
9	Exposure to pornography in childhood	8.13	1.3	9	- Highly prevalent, affecting 1 in 12 children. Severity varies with the evolving capacity of children as they age. - Suggestion to rephrase as "exposure to unwanted sexually explicit content in childhood". - Recommendation to combine with **6** under a broader category of "unwanted content".
10	Illegal transactions (including drugs)	7.83	1.27	9	- Limited data on prevalence, but self-reporting indicates use of social media for these purposes.
11	Loss of control, addiction, excessive use	7.74	1.92	9	- "Addiction" is a contested and possibly stigmatizing term. - Recommendation to combine with **13**, **18**.
12	Financial harms: underage access to illegal gambling or excessive spend on in-app transactions	7.57	1.38	9	- Limited data on prevalence, but anecdotal evidence of teenage access to online gambling via social media. - Not youth-specific; salience to youth users must be clearly articulated for risk assessment purposes.
13	Displacement of other beneficial activities (including sleep, exercise and in-person social activities)	7.48	1.78	9	- Mechanism, not a harm. - Recommendation to combine with **11**, **18**.
14	Misogyny, racism, hate speech	7.4	1.63	9	- Mechanism, not a harm. - Rooted in societal issues that predate and transcend social media.
15	Communities/content that promote dangerous challenges, or unsafe or unhealthy products	7.32	1.74	9	- Potentially very high-severity, including known youth deaths. - Recommendation to combine with **3**, **5**.
16	Fraud (including identity theft, impersonation, scams)	6.87	1.87	6	- Wide-ranging severity. - Not youth-specific; would require clear articulation of youth characteristics that could exacerbate general risk. - Exclusion from risk assessments could send the wrong signal.
17	Upward social comparison	6.26	2.19	9	- Mechanism, not a harm. - High prevalence. - Identified by youth users as a key issue.
18	Psychological impacts, including depression, sadness, anxiety, loneliness, lower positive well-being indicators, such as happiness, self-esteem	6.09	2.34	8	- High level of heterogeneity, with different hypothesized mechanisms, complicates inclusion in risk assessments. - Recommendation to combine with **11**, **13**.
19	Misinformation and disinformation	5.83	2.32	7	- Mechanism, not a harm. - Not youth-specific. - Rooted in societal issues beyond social media.
20	Infringement of Child Rights: over-limiting child access to information	5.74	2.22	5	- Suggested as an additional category following preliminary survey. - Not discussed in detail due to time constraints.
21	Algorithmic biases and risks (including recommendations of problematic content or discrimination in decision-making)	553	2.5	8	- Not well defined. - Mechanism, not a harm; should be considered in relation to other harms. - A tool that can exacerbate harms, but also underpins many mitigation strategies.
22	Parent use of social media and related stress/displacement of social interactions	5.27	2.56	5, 9	- Mechanism, not a harm. - Worth addressing in parent-targeted materials.

Table 1: Integrated Harm Framework, with final voting statistics and key points of discussion

Across topics, a few overall themes emerged—while these were also not necessarily consensus points, they did occur consistently in conversations across multiple harms and influenced attendee scoring. Some of these themes include:

1. **Prevalence vs. severity:** The prevalence (how often a harm occurs) and severity (how impactful a harm is when it occurs) vary significantly by harm. Some harms are high severity, low prevalence, while others are low severity, high prevalence. Some participants emphasized the importance of building a shared understanding of prevalence and severity for each harm and noted that the strength of the evidence base varies by harm.
2. **Mechanisms vs. outcomes:** Some harms were more mechanistic (e.g. 'social comparison') without a specified outcome, while others were specified outcomes without a specific mechanism (e.g. 'psychological impacts, including depression'). Some participants argued against mechanism-only harms; they were more in favor of ones that included outcomes and suggested amending the labeling of some mechanism-only harms to specify an outcome (e.g. 'social comparison that results in anxiety or reduced self-esteem').
3. **Actionability:** In specifying harms for inclusion, some participants emphasized the importance of considering how the specification will be used, and favored harms that were specific enough to be measured and mitigated. Whether a given phenomenon is a harm or a mechanism was noted as a dimension that affects actionability. Another dimension was whether the specific harms conceptually map into how platforms currently approach mitigating issues in practice, with stronger alignment being more actionable.
4. **Youth-specific vs. general harms:** Some harms are generally applicable to all age groups, whereas others may be differentially harmful to youth (e.g. 'exposure to unwanted pornography in childhood'). Some participants argued for more of a focus on the latter; others suggested prioritizing by severity and prevalence for youth, regardless of whether something is uniquely harmful for youth.
5. **Societal issues vs. social-media-only issues**: Some harms have pre-existing offline components (for example, bullying occurs in schools), and in some cases online components of those harms may only be part of broader societal issues. Some participants emphasized the importance of considering these harms as part of a broader system for both assessment (e.g. trying to understand relative prevalence and severity across channels) and mitigation (e.g. considering both online and offline resources in mitigations). Some also suggested a heightened priority for the harms that are differentially over-represented online.

Measurement and Metrics

In the second part of the workshop, three breakout groups convened (one for threats to safety/criminal activity; one for health & wellbeing; and one for other content-driven harms and other harms). Each group discussed approaches to measure specific harms in their category. Some shared key points of discussion are summarized below—again, these were not necessarily consensus points.

Recommended Approach

Start with the questions you are trying to answer, and then select the best measurement approach and metrics for those questions.

Risk assessment questions:

1. Prevalence ('How frequently does the harm occur?')
2. Severity ('When the harm occurs, how problematic is it?')

Risk mitigation questions:

1. Effectiveness ('How much of the risk of the harm does the mitigation reduce?')
2. Feasibility ('How reasonable is the mitigation strategy? Are there other trade-offs, including costs, privacy, free speech, or loss of benefits?')

High-risk groups: In addition to the population as a whole, it is useful to answer these questions for specific high-risk groups such as LGBTQ+, under-resourced youth, and youth with significantly more adverse childhood experience, since risk of harms may not be evenly distributed.

Public vs. private contexts: Measurement may need different approaches for public or semi-public content feeds vs. encrypted or private messages and groups. The latter introduces measurement challenges.

Metric Types

1. **Content or behavior metrics.**
    a. Metrics such as screen time, frequency of interactions, and engagement with specific types of content.
    b. For content-based metrics, analysis of random samples can help estimate prevalence of content. Except for cases where the content is the harm, such as Child Sexual Abuse Material (CSAM) and Non-Consensual Intimate Image (NCII) abuse, the metrics are only proxies for the actual harm.
    c. Availability: These data are typically readily available to platforms with low latency.

2. **User feedback and reports.**
    a. Metrics include self-reported experiences from representative samples of the population; in-platform feedback reporting mechanisms, and reports through other channels (e.g., law enforcement).
    b. User report data, including both via representative surveys and via user reporting through customer service channels, can capture the personal impact of harms, including those that may not be visible through automated detection. Users may lack self-awareness in the case of some specific harms (e.g., body image content; displacement of beneficial activities). User reporting is likely to be limited by perceived barriers to reporting, including perceptions that reporting is ineffective.
    c. Availability: These data are sometimes available to platforms but surveys of representative samples of the population are obtainable externally. Typically higher latency than content or behavior metrics.

3. **Validated offline outcomes (e.g. health outcomes; some are self-reported)**
    a. Metrics include validated measures of anxiety, depression, etc.
    b. These metrics are the most definitive but also the hardest to collect at scale.
    c. Availability: These data are sometimes available to platforms, although policy and legal restrictions mean that platforms do not have these widely available. These are obtainable externally. Typically higher latency than content or behavior metrics.

Measurement challenges

1. **Lack of standardized metrics:** Variability in definitions and tools complicates cross-platform and cross-population comparisons. Surveys trying to measure the same thing also often have different questions and do not always use questions that are validated in the academic literature.

2. **Lack of data on impact of platform mitigations:** understanding the impact of platform risk mitigations (e.g. product features offering greater control, algorithmic ranking changes) typically requires access to interventional data, which are not publicly available (though they may be available to platforms and in some cases regulators). This makes it hard to assess their impact.

3. **Lack of data on prevalence of online vs. offline harms:** Some of these harms also occur offline and have complex offline/online interactions. Understanding relative prevalence of online vs. offline harms can inform prioritization and approaches to solutions. In some cases these data exist (e.g. CDC bullying surveys), but in others they do not.

4. **Discrepancies in detection algorithms:** Automated detection systems used by platforms for content moderation or identifying problematic user behavior vary widely in accuracy. This can result in underreporting or overreporting of harmful content depending on the platform's technological sophistication and biases in its algorithms.

5. **Subjectivity/varied impact across individuals:** Many social media harms are difficult to quantify because they depend on users' subjective experiences. Differences in cultural context can be a complicating factor: for instance, what is considered misogynistic may vary. Psychological impacts such as depression, anxiety, and social comparison are specific to each individual. Different people respond differently to the same social media content, making it challenging to capture consistent patterns of harm.

6. **Variation in quality of training data by language:** In the context of the EU, the small sample sizes for some national languages in the bloc make the training of models to detect certain kinds of harmful content unfeasible.

Mitigation

In the third part of the workshop, the same three breakout groups reconvened. Each group discussed approaches to mitigate specific harms in their category. Some shared key points of discussion are summarized below—again, these were not necessarily consensus points.

General Considerations

1. **Tailoring & youth involvement:** In some cases, mitigation strategies targeted at youth-specific online risks should be tailored for this demographic, as they may differ from more generalized strategies. Youth voices should be included in the development of mitigation strategies.
2. **Realistic aims:** Mitigation strategies should balance idealism with pragmatism; reduction in the prevalence of harms is a positive, even if full eradication is not achievable. It is consequently necessary to establish an acceptable level of residual risk for each harm.
3. **Encryption:** Multiple harms may be prevalent in encrypted messaging channels, which require unique strategies to measure and mitigate. Platforms may be able to access content if a report is made, but ease of reporting content and practices for reviewing reported content across platforms varies. In the case of some harms, like illegal transactions, it is also unlikely that participants will report, given that reports would constitute self-incrimination. Scanning mechanisms have been proposed but also introduce privacy risk if they are repurposed beyond just illegal content.
4. **Industry cooperation:** Cross-platform and cross-sector collaboration is vital, especially in cases of CSAM and other severe harms. In some cases the largest drivers for mitigation are within the platforms; in other cases, they may be outside of the platforms or at a minimum require strong cooperation (e.g. via law enforcement, health and medical authorities, or via schools or parents/families).

Mitigation Strategies

1. **Age verification:**
    - Crucial for mitigation strategies tied to age, although the strength of verification required depends on the mitigation strategy. For example, Android and iOS already offer 'child accounts' that parents can set up when a child receives a device; for some use cases, this lightweight 'verification' may be sufficient, though data is lacking on adoption rates. More stringent verification methods such as requiring official identification may offer stronger protections, but also introduce privacy and accessibility trade-offs.
    - Platforms should strengthen default privacy measures for youth users, with restrictions on features including private messaging and profile recommendations between adult and minor users. It may be worthwhile to separate discussion of age verification strategies from

discussion about default measures and functionality for minor accounts, since these can be implemented in parallel—and age verification is in service of the mitigation strategies.
    - Facial recognition and other verification methods have privacy implications.
    - Behavioral monitoring is used widely in the EU for age estimation, but estimate-based approaches risk 'false positive' identification of minors, potentially infringing on the rights of adults.
2. **Education:**
    - Develop digital literacy resources, including workshops and school curriculum components, to empower youth to navigate social media safely and make informed decisions about usage. Evaluate impact of these educational resources, in both on-platform and outside-of-platform settings.
    - Educational approaches are important for parents, teachers, communities as well as minors/adolescents themselves. Provide families with accessible, evidence-based tools to manage digital use. Focus on fostering open conversations about risks and benefits.
    - Community-driven and platform-driven digital education have the potential to be synergistic. Work by parents, families, and educators to mentor youth, along with platform design that is transparent and youth-centered, will help young users be the most informed and intentional in their relationship with technology.
3. **Content Moderation:**
    - There is a growing realization that content moderation is important, but that there is an additional opportunity to address the 'upstream' causes of the spread of harmful content. A key issue is the algorithmic promotion of harmful content, driven by ranking approaches centered around engagement.
    - Only a subset of the harms (e.g. CSAM, eating disorders) are content-only harms. Many are tied to behaviors or experiences (e.g. harassment; solicitation). Content moderation is not necessarily the best or only approach to more complex harms.
4. **Algorithmic interventions:**
    - Virality-dampening can systematically reduce the spread of harmful content. While effective, this approach can have unintended consequences, such as restricting the reach of social movements. Algorithms should be designed to prioritize high-quality content rather than attention-grabbing content—but there is a consequent need for consistent metrics determining what is high-quality.
5. **Youth-Interests-First Designs:**
    - A related approach is developing features that encourage users to be more intentional in their usage, including developing features and mechanisms to encourage users to align their use and their own intentions and interest and facilitate self-reflection and self-regulation, such as reminder of users' goals, pop-up such as 'Think before you post' messages, and periodic reminders of community rules.

Key Insights and Recommendations

In the last part of the workshop, the full group reconvened. Participants were asked to individually share key insights and recommendations that they had for platforms, policymakers, researchers, and parents/caregivers regarding all the possible harms the group discussed. The Integrated Harm Framework (IHF) emerged as a unifying model to guide these efforts, offering a shared foundation for developing evidence-based and actionable strategies. Some key points of discussion are summarized below—again, these were not necessarily consensus points.

For Platforms:

- **Enhance transparency and data sharing**
  - Identify the key prevalence and mitigation impact questions that uniquely require access to platform data, and work to provide this access to regulators and independent researchers.
  - Publish regular reports detailing their best understanding of prevalence, severity, and their most impactful mitigations for specific harms.
- **Refine algorithms and platform design**
  - Prioritize content quality over engagement metrics.
  - Employ virality-dampening measures to limit the spread of harmful content while safeguarding free speech.
  - Offer customizable content controls for users to increase agency, enabling users to filter or reduce exposure to potentially harmful material.
  - Reduce features that may exacerbate extended and less intentional usage (e.g. push notifications, infinite scroll). Make it easier to toggle particular platform features on or off, and consider having certain features off by default for youth users.
  - Design digital spaces for more intentional and valuable usage.
- **Provide tools and resources**
  - Invest in tools that give people agency to control their experiences and self-remediate. For example, tools that allow people to keep their accounts ‘private by default’ or to restrict or limit who can interact with them.
  - Use on-platform, in-context tools to promote awareness of risks and provide access to resources, including those off-platform. This can include mental health resources, Non-Consensual Intimate Image (NCII) abuse and other helplines, and in some cases, connections to victim services or law enforcement, ideally delivered to the people who report these issues.
  - Conduct usability testing and user experience research on parental controls (including on adoption rates, parent and child satisfaction, and overall effectiveness) and share the results publicly.
  - Improve parental control tools so that they are easy to use and accessible to parents with all levels of digital literacy.

For Policymakers:

- **Establish consistent, global standards**
  - Develop regulations to ensure proportionate safety measures across platforms, reducing disparities in harm mitigation efforts between countries and platforms while reflecting differences in organization size/capacity and risk surface.
  - Focus on creating youth-specific guidelines, such as age-appropriate content moderation, data privacy, design features, and approaches to age verification.
  - Share knowledge, metrics, and methods between regulatory bodies across jurisdictions.
- **Enhance transparency and data sharing**
  - Identify the key prevalence and mitigation impact questions that uniquely require access to platform data, and enforce requirements to provide this access to both regulators and vetted, independent researchers.
  - Implement safeguards to ensure privacy protections and ethical use of shared data by researchers.
  - Identify key prevalence and mitigation impact questions that can be answered without unique access to platform data, and ensure that efforts to answer these are funded. For example, standardized, representative consented surveys of youth on harms and benefits.
  - Require or strongly encourage independent audits of reports provided by platforms to regulators and to the public.
- **Encourage industry collaboration**
  - Facilitate and incentivize cross-platform cooperation to tackle shared challenges. Some entities already exist for CSAM and terrorism; they could be expanded for other harms that are tied to activity that could be illegal (e.g. scams/fraud, some forms of harassment/stalking). Provide funding for joint initiatives that promote innovation in harm mitigation, particularly for harms such as sextortion that are multi-platform in nature. Funding could come from platforms, either voluntarily or through regulation, or from existing appropriations processes.
- **Fund and modernize approaches to illegal activity online**
  - Ensure that organizations like the National Center for Missing and Exploited Children (NCMEC) and the Internet Crimes Against Children (ICAC) Task Force, which handle enforcement of some crimes that occur online, are adequately resourced, with robust technical systems.
  - Ensure that enforcement and support for victims of illegal activity online is strong and that there are no gaps when victims go to law enforcement for assistance. Provide guidance to victims and law enforcement on how to document illegal activity online.
- **Balance rights and responsibilities**
  - Address the tension between privacy, free speech, and safety by fostering public dialogue and transparent policymaking that includes youth and family perspectives.
  - Ensure that regulatory measures do not disproportionately affect marginalized groups or limit access to beneficial content.
- **Consider the role of schools and families**
  - Schools can play a meaningful role in educating children and parents about policies on phone use, as some harms (e.g. bullying and harassment) extend into offline or in-school interactions.
  - Consider regulation to make parental control functionality mandatory for all platforms that could cause harm to minors, and require regular usability testing and transparency reporting about these features to assure they are accessible and functional for families from all backgrounds.
  - Avoid placing sole responsibility on parents to manage their children's digital lives. Provide resources and systems that enable effective engagement and communication with children and teens, without adding undue burden and blame.
- **Youth centered approach**
  - Policies should be centred on, and guided by, the voices of children and young people, and their life experiences.

For Researchers:

- **Focus on practical, actionable research**
    - Prioritize studies that directly inform policy and platform design, emphasizing measurable outcomes and clear recommendations.
    - Do not overlook areas under-addressed by existing research, including but not limited to the impact of specific platform design changes on youth wellbeing, parental social media use, youth access to online gambling, use of social media for illegal transactions, and algorithmic bias detection.
- **Bridge the gap between science, policy, and the public**
    - Translate academic findings into accessible, policy-relevant language to inform decision-makers and the public.
    - Conduct more intervention studies to understand what works best in experimental settings and inform regulations.
    - Collaborate with interdisciplinary teams and platform designers to find a balance between user interests, business models, health and well-being. Engage in discussions with policymakers and seek out opportunities to propose specific regulations.
    - Collaborate internationally, especially with researchers inside and outside of jurisdictions that offer superior researcher data access, such as the EU, such that lack of data access in other jurisdictions does not impede the progress of research that could be partially or fully generalizable across geographies.
- **Engage with youth**
    - Involve young people in participatory research and design to represent their perspectives on risks and solutions.

For supporting Parents and Families:

Discussion generally centered around how to support parents and families as opposed to specific steps that parents themselves can take. An overarching point of consensus was that guidance and information for parents should not be prescriptive or 'one-size-fits-all', but should be flexible and take into account the diversity in family circumstances and relationships to tech.

- Provide reliable, accessible information on how platforms work and what safety settings are available to help parents navigate digital safety challenges. Focus on practical guidance that reduces fatigue and builds confidence.
- Parents can encourage the involvement of school-aged children and teens in establishing boundaries around social media use collaboratively to foster their sense of agency and create a balanced digital environment at home.
- Promote community engagement with others facing similar challenges. Shared experiences and collective action can normalize the learning curve in digital parenting.

Acknowledgments

We extend our sincere gratitude to the attendees of the Youth Safety and Digital Wellbeing workshop for their invaluable contributions. The collective knowledge and thoughtful engagement of each participant greatly enriched our understanding and were essential to the learnings compiled in this final report.

Disclosures

The work in this report was funded by Stanford University.

Attendees of the workshop are listed below. The opinions expressed in the workshop were those of the individuals, not the organizations they are affiliated with. The document represents some of the main points discussed at the workshop, but these were not necessarily consensus points.

Workshop Attendees

Eleni Linos, MD, DrPH, Stanford Center for Digital Health
Jeff Hancock, PhD, Stanford Social Media Lab
Michael Avanti Lopez, JD, Stanford Center for Digital Health
Ravi Iyer, PhD, USC Marshall School Neely Center
David Harris, MS, University of California, Berkeley
Vicki Harrison, MSW, Stanford Center for Youth Mental Health & Wellbeing
Angela Yuson Lee, MA, Stanford Social Media Lab
Jenny Radesky, MD, AAP Center of Excellence on Social Media and Youth Mental Health
Alexis Hiniker, PhD, University of Washington
Alissa Cooper, MS, DrPhil, Knight-Georgetown Institute
Anja Stevic, PhD, Stanford Social Media Lab
Sunny Liu, PhD, Stanford Social Media Lab
Anna Lembke, MD, Stanford University
Roberta Katz, PhD, Center for Advanced Study in the Behavioral Sciences at Stanford
Jennifer Heifferon, MAT, California Partners Project
David Sullivan, MA, Digital Trust and Safety Partnership
Yvette Renteria, M.Ed, Common Sense Media
Kang-Xing Jin, Stanford Center for Digital Health, Former Meta
Alicia Blum-Ross, PhD, TikTok
Dave Willner, Zentropi, Former OpenAI
Joanna Smolinska, European Union Office in San Francisco
Gerard de Graaf, European Union Office in San Francisco
Giulia Geneletti, European Union Office in San Francisco
Annika Ostergren, European Commission
Martin Harris Hess, European Commission
Mariesa Nicholas, eSafety Commissioner, Australia

https://cdh.stanford.edu
DigitalHealth@Stanford.edu
X: @StanfordCDH

https://sml.stanford.edu
team@ssml.stanford.edu
X: @StanfordSML