Consultation: Draft transparency reporting guidance Ofcom

4 October 2024

Response from the Center for Countering Digital Hate (CCDH)

Introduction

The Center for Countering Digital Hate (CCDH) is an international non-profit research and advocacy organisation with offices in Washington D.C., Brussels and London. CCDH conducts independent studies, campaigns for social media reform, and advises lawmakers and regulators on the basis of our research.

The following submission is drawn from CCDH's research into online harms. Particularly relevant to this consultation is the CCDH STAR Framework, our blueprint for social media reform and transparency, and our practical experience with the challenges of researching opaque social media platforms and search services.[1] CCDH has been targeted with legal action over our public-interest research, giving us unique insight into the lengths some platforms will go to avoid transparency and accountability.[2]

CCDH has supported the UK Online Safety Act since its inception.[3] Here we offer advice on the implementation of the Act's transparency reporting duties and discuss what information Ofcom should detail in its annual transparency report to be most useful to CCDH and the wider sector.

CCDH feedback on Ofcom's guidance to categorised service providers

CCDH welcomes the overall approach taken by Ofcom and agrees that transparency is an indispensable tool for raising safety standards and imbedding safety in product design and

[1] "Building a safe and accountable internet: CCDH's refreshed STAR Framework" CCDH, Sep 2024, https://counterhate.com/wp-content/uploads/2024/09/CCDH.STAR-Framework.Report-FINAL.pdf

[2] "Elon Musk vs. CCDH: nonprofit wins dismissal of 'baseless and intimidatory' lawsuit brought by the world's richest man" CCDH Blog, 25 Mar 2024, https://counterhate.com/blog/elon-musk-vs-ccdh-nonprofit- wins-dismissal-of-baseless-and-intimidatory-lawsuit/

[3] see Imran Ahmed, oral evidence to Draft Online Safety Bill (Joint Committee), 9 Sep 2021, https://committees.parliament.uk/oralevidence/2693/pdf/ (transcript); https://www.youtube.com/watch?v=X5JdQhAjnVEClist=PLEcb8pRWfsxUZBRoCyz2-RAyb5c6HrnSBCindex=2 (recording)

business decisions. However, we are concerned that the current proposals leave too much scope for influence by the industry and suggest clarifying the industry engagement proposals. We are also concerned that Ofcom is improperly applying proportionality considerations that are out of place in this area of the regulation given its (anticipated) application to only the largest platforms, which are best resourced to comply with regulation.

1. Avoid overreliance on industry reporting metrics

CCDH research has found that the social media industry's own reporting metrics produce incomplete or misleading insights into online harms.[4] In public statements and self-published transparency reports, platforms represent these metrics as sufficient or effective at addressing online harms, but internal company records reveal that services are aware of the limitations of these reporting metrics, and in some cases use them actively to obscure reality.[5] The clearest example of this is the widespread utilisation of "prevalence". These metrics, created by industry to judge itself against, have become the norm for many of the services in scope of the transparency reporting duties (subject to the Secretary of State's categorisation decision). Schedule 8 gives Ofcom sufficient scope to require information relating to illegal content and harms to children, but given this widespread industry orientation towards incomplete or vague reporting metrics, there is a risk that unless discredited industry metrics are explicitly counteracted, these reporting methods will become default tools in responding to regulatory requirements. Ofcom should be setting out its own metrics, rather than conforming to a precedent set by industry. CCDH would draw Ofcom's attention to prior examples which should be designed against in the transparency notices sent to categorised platforms:

> **Prevalence** – Prevalence is a dominant metric used by social media companies for disclosures about their safety programmes. The "prevalence" of content which violates a platform's community standards or terms of service is estimated using a sample of content on the platform, assessing it for violations and labelling it if so. From the result of this assessment, a platform estimates how common violative content is on its service as a whole. However, as pointed out by CCDH and Meta

[4] "Fact-checking TikTok's claims on Antisemitism" CCDH Blog, 6 Nov 2023, https://counterhate.com/blog/fact-checking-tiktoks-claims-on-antisemitism/

[5] Unredacted federal complaint filed by 33 attorney generals against Meta Platforms, Inc. https://oag.ca.gov/system/files/attachments/press-docs/Less-redacted%20complaint%20-%20released.pdf
"Fewer than 1% of parents use social media tools to monitor their children's accounts, tech companies say" NBC News, 29 March 2024, https://www.nbcnews.com/tech/socialmedia/fewer-1-parents-use-social-media-tools-monitorchildrens-accounts-tech-rcna145592

whistleblowers like Arturo Bejar, prevalence assessments are drawn only from content that a platform can identify and label, which may not be truly representative of violative content on the whole.[6] By comparing these identified instances against the massive denominator of all content on the service, prevalence functions to obscure more than illuminate. Further, social media is an individualised experience, in that algorithms and recommender systems tailor experience to user data, meaning a metric based on the overall proportion of violative content misses how users experience harm (or encounter illegal content) and obscures critical facts about the safety of the platform. To avoid bedding in this current practice, CCDH suggests that the draft guidance for service providers be amended to reduce or clarify references to this discredited industry reporting metric (example, pg 7).

Discoverability is a more optimal metric. Discoverability is the ratio of violative content to the relevant content area. For example, to assess the rate of eating disorder content, this metric would assess the scale of the content that contravenes OSA rules against the total number of views on all content relating to eating and dieting. In this way, it does not disguise the true scale by inflating the denominator.

AI moderation – Platforms exaggerate the effectiveness of their AI moderation systems. In public pronouncements and voluntary transparency reports, platforms claim these moderation tools identify and remove the vast majority of violative content.[7] But internal company communications reveal a starkly different reality. In 2021, one of Meta's senior research scientists estimated that the company's AI tools caught content that was responsible for just 2% of all the views of hate speech on the platform, and a separate team concluded that the company's automated systems removed content that generated just 3% to 5% of views of hate speech.[8] While Meta employees were internally calling attention to the limits of AI content moderation, senior leadership was publicly claiming that its AI tools proactively detected 98% of all violating content.[9] In another example, Meta's quarterly

[6] "Recommendations for Regulators" Arturo Bejar, accessed 3 Oct 2024, https://docs.google.com/document/d/14jVJ_XSwv-bgwgawMRC37ZVGObvUEYafzwRbEb_aA5U/edit?tab=t.0#heading=h.hn0ilpna8a83

[7] "Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts." Wall Street Journal, 17 Oct 2021, https://www.wsj.com/articles/facebookai-enforce-rules-engineers-doubtful-artificialintelligence-11634338184

[8] ibid.

[9] ibid.

transparency reports cited a higher detection rate for child abuse content than proved to be true following independent investigations.[10]

Classification decisions – Dubious platform choices over how to classify users and moderation decisions are at the root of many misleading public pronouncements and voluntary transparency reports. As revealed by Clean Up The Internet, public claims from Twitter about online racism following the 2021 Euros were based upon a suspect classification decision over what constituted an anonymous account.[11] Another example is the vagueness of the term "actions" that platforms like Meta and TikTok use to record content moderation processes, as it can include anything from minor pop-up warnings to major interventions like alerting law enforcement authorities.[12] The widespread utilisation of these nonrepresentative transparency metrics and dubious classification decisions should be counteracted in Ofcom's transparency reporting guidance.

CCDH believes that transparency duties must require categorised services to report different metrics than those they have historically used, such as the discredited metric of "prevalence" and nebulous term "actions". In the following section, we suggest metrics and the types of analysis and insight that would best assist in our work.

2. Clarify "consideration" to avoid creating an influencing pathway

The draft guidance describes other factors that will affect Ofcom's design of transparency notices. In paragraph 3.26 (pg 11), Ofcom says it will first consider "whether the information has already been provided or published", taking note of information included in services' voluntary transparency reports, and as available via reports to other regulatory regimes. The extent to which these factors will be considered must be clarified to avoid creating an undue influence pathway for categorised services.

[10] "Facebook blames glitch after huge drop in child abuse image takedowns", The Daily Telegraph, 19 May 2021, https://www.telegraph.co.uk/technology/2021/05/19/facebook-blames-glitch-huge-drop-child-abuse-image-takedowns/

[11] "Combatting online racist abuse: an update following the Euros" Twitter UK, 10 Aug 2021, https://blog.x.com/en_gb/topics/company/2020/combatting-online-racist-abuse-an-update-following-the-euros

"Twitter's anonymity claims appear to rely on classifying Mickey Mouse accounts as "not anonymous"" Clean Up the Internet, 10 Dec 2021, https://www.cleanuptheinternet.org.uk/post/twitter-s-anonymity-claims-appear-to-rely-on-classifying-mickey-mouse-accounts-as-not-anonymous

[12] "Transparency is essential for effective social media regulation" Brookings Institute, 1 Nov 2022, https://www.brookings.edu/articles/transparency-is-essential-for-effective-social-media-regulation/

As described above, voluntary transparency reports have been shown to contain misleading claims. While it is reasonable to consider what information services already include in their voluntary transparency reports, given the evidence of misleading and incomplete information, Ofcom must not consider voluntary transparency reports and pronouncements as a substitute for information requested via transparency notices. To strengthen 3.26, CCDH believes that "consideration" should be narrowly defined as "taking note of", rather than more broadly as a pathway to adjustment.

It is worth noting here that information placed in the public domain via other regulatory regimes has been variable, and should therefore also be narrowly considered for the purposes of Ofcom's transparency notices. For example, TikTok launched a Research API in 2023 to meet the requirements of the European Union Digital Services Act.[13] Researchers used the API to collect data in advance of the European Parliamentary Elections in 2024. However, when cross-checking data obtained through via the API, researchers noticed significant deviations between this data and data visible on TikTok's application or website.[14] The point here is that Ofcom's transparency notices should not accept at face value information in the public domain via other regulatory regimes, as in this example that information proved to be inaccurate.

3. Transparency over any alterations made during the engagement process

Ofcom describes how it will engage with categorised services during the drafting of transparency notices (paragraphs 4.13 and 4.14). This engagement is not required by the Online Safety Act, but has the stated aim of allowing clarifications and feedback on notices before being formally issued. CCDH believes that this pre-issuing engagement presents a risk of undue influencing by industry. These sections must be re-drafted to ensure that platforms do not use these conversations as an opportunity to water down the requirements of their transparency notice. After this clarification, Ofcom should also commit to transparency over any alterations to a notice made as a result of these pre- issuing discussions.

[13] "Expanding TikTok's Research API and Commercial Content Library" TikTok, Jul 2023, https://newsroom.tiktok.com/en-eu/expanding-tiktoks-research-api-and-commercial-content-library

[14] "Researcher Data Access Under the DSA: Lessons from TikTok's API Issues During the 2024 European Elections" Tech Policy Press, 24 Sep 2024, https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections/

4. “Proportionality” assessments

Paragraph 3.20 discusses the principle of proportionality, saying that Ofcom will have the relevance, appropriateness, and proportionality of transparency requests in mind when utilising its statutory powers. But CCDH is concerned that Ofcom is generally interpreting “proportionality” in this consultation in terms of associated costs, as it did in earlier consultations on illegal content and harms to children. CCDH and others raised concerns about this narrow interpretation in earlier consultation responses, but these concerns are even greater in this area of the regulations given that transparency reporting only applies to the largest services with the greatest level of financial resource.[15]

There is also evidence of platforms claiming costs that cannot be verified, but have been judged suspect in legal proceedings. For example, in X’s lawsuit against CCDH, the company claimed that our research had resulted in significant costs to their business for computing repair and server processing.[16] In his dismissal of the suit against CCDH, the presiding judge was unconvinced by X’s cost allegations, saying that small-scale, non-commercial research such as CCDH was conducting could not plausibly have cost the sum that X alleged.[17] In summary, Ofcom must treat any claim for costs by categorised services that cannot be independently verified with skepticism and not allow such claims to influence their proportionality assessment.

CCDH feedback on Ofcom’s transparency report

CCDH launched an updated version of our STAR Framework in September 2024. The STAR Framework is a globally applicable blueprint for regulating social mediaand outlines the transparency metrics and information necessary to ensure accountability and truly independent oversight.[18] It is contained as an annex to this submission. In reference to

[15] see CCDH’s response to Ofcom’s illegal harms consultation: https://counterhate.com/research/ccdhs-ofcom-illegal-harms-consultation/

[16] see X Corp v. Center for Countering Digital Hate, United States District Court, Northern District of California, 25 Mar 2024, https://casetext.com/case/x-corp-v-ctr-for-countering-dig-hate

[17] ibid. See reference s.41: ”It is not plausible that this small-scale, non-commercial scraping would prompt X Corp. to divert ’dozens, if not over a hundred personnel hours across disciplines,’ see Tr. of 2/29/24 Hearing at 8:7-11, of resources toward the repair of X Corp.'s systems.”

[18] “Building a safe and accountable internet: CCDH’s refreshed STAR Framework” CCDH, Sep 2024, https://counterhate.com/wp-content/uploads/2024/09/CCDH.STAR-Framework.Report-FINAL.pdf

STAR, CCDH urges Ofcom to include the following information in their transparency reports:

1. Insights on differences between public and private transparency reporting

Ofcom should include analysis of areas where information received from platforms in their transparency reports aligns or does not align with information contained in their voluntary transparency reports and public pronouncements. This is a critical assessment to include, as it will encourage platforms to raise their standards of public truthfulness and start unlearning their instinct to mislead and obscure in public pronouncements (as extensively evidenced above). An example of the current mismatch between voluntary transparency reporting and transparency data reported under regulatory requirements was done by the Molly Rose Foundation, which analysed 12 million content moderation decisions by major tech platforms recorded under the rules of the EU Digital Services Act. They found that 98% of all moderation decisions on suicide and self-harm content were taken by just two platforms, TikTok and Pinterest, and that there were significant and seemingly irreconcilable differences between the number of moderation decisions Meta platforms reported in their DSA filings and the decisions they claim to have taken in their voluntary transparency reports.[19] Ofcom should similarly assess these differences and include that assessment in its transparency report.

2. Insights on how harmful content is experienced by users

Ofcom writes that the goal of its transparency reports are to "empower UK users with relevant and accurate information about risks and safety outcomes on services" (paragraph 5.3). To meet this goal, it is critical that Ofcom translates the information it receives in provider reports into insights on how these matters are experienced by users. Platforms have often responded to research evidencing online harm by saying that the findings are "not reflective of the experience or viewing habits of real-life users on the app" (see TikTok's response to CCDH's 2022 report Deadly By Design).[20] By this cynical response we are to gather that user experience is only truly knowable by the platform and that countless individuals' first-hand experiences, or replications of it by researchers, are

[19] "How effectively do social networks moderate suicide and self-harm content?" Molly Rose Foundation, Aug 2024, https://mollyrosefoundation.org/wp-content/uploads/2024/08/DSA_Transparency_report_MRF.pdf

[20] "TikTok self-harm study results 'every parent's nightmare'" The Guardian, 15 Dec 2022, https://www.theguardian.com/technology/2022/dec/15/tiktok-self-harm-study-results-every-parents-nightmare

not representative. Thus, Ofcom must test the truthfulness of these platform claims about user experience and address this information asymmetry in its own transparency reports.

#3. Counteracting measures to shut down oversight with Ofcom transparency report

Data and information about social media content and a product's design are some of the most valuable resources for understanding online harms, but in recent years social media companies have restricted or eliminated the tools researchers use to access that information.[21] In 2023, X cut off researchers' ability to access data via its API and began charging $42,000 per month for the previously free service, disrupting hundreds of independent research projects that relied on the data.[22] X's new leadership has gone on to sue independent researchers, including CCDH.[23] A survey of independent researchers found that a majority of respondents fear being sued by X over their findings or use of data.[24] In 2021, Meta abruptly cut off researcher's access to transparency tools and in 2023 shut down shut down CrowdTangle, a platform monitoring tool that was used by many independent researchers.[25] It is in this context that Ofcom will be publishing its transparency report. The draft guidance makes clear that Ofcom recognises the role its transparency reports will play for independent researchers, but the draft documents appear to be crafted with industry top-of-mind. CCDH urges Ofcom to recognise the significant information asymmetry between the major social media platforms and those who wish to hold them to account. To meet the stated ambitions for transparency reporting outcomes, Ofcom must counteract efforts by industry to reduce access for researchers with contextually rich information and data in its transparency reports.

---

[21] "Meta to Replace Widely Used Data Tool—and Largely Cut Off Reporter Access". The Wall Street Journal, March 14, 2024, https://www.wsj.com/tech/meta-to-replace-widelyused-data-tooland-largely-cut-off-reporter-access43fc3f9d

[22] " Twitter just closed the book on academic research". The Verge, May 31, 2023. https://www.theverge. com/2023/5/31/23739084/twitter-elon-musk-api-policychilling-academic-research

[23] " Musk's Meltdown Timeline: CCDH against Elon Musk's attacks on independent research" CCDH, 1 Sep 2023, https://counterhate.com/blog/musks-meltdown-timeline-elon-musks-attacks-on-independent-research/

[24] Exclusive: Elon Musk's X restructuring curtails disinformation research, spurs legal fears. Reuters. November 6, 2023. https://www.reuters.com/technology/elon-musks-xrestructuring-curtails-disinformation-research-spurslegal-fears-2023-11-06/

[25] "Facebook is obstructing our work on disinformation. Other researchers could be next". The Guardian. August 14, 2021. https://www.theguardian.com/technology/2021/aug/14/ facebook-research-disinformation-politics

"Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn't as Transparent or Accessible". Columbia Journalism Review. July 9, 2024. https://www.cjr.org/tow_center/metais-getting-rid-of-crowdtangle.php

Conclusion

CCDH thanks Ofcom for the opportunity to respond to this consultation. By making clear that discredited industry reporting metrics are not acceptable, shutting down unintended pathways for industry influence, and making more explicit commitments to harm reduction as an intended outcome of the process, CCDH believes transparency reporting will go a long way to making the UK the safest place in the world to be online.

Annex

Full PDF of STAR Framework

Center for Countering Digital Hate

BUILDING A SAFE AND ACCOUNTABLE INTERNET:

CCDH'S REFRESHED STAR FRAMEWORK

The Center for Countering Digital Hate works to stop the spread of online hate and disinformation through innovative research, public campaigns and policy advocacy.

Our mission is to protect human rights and civil liberties online.

Social media platforms have changed the way we communicate, build and maintain relationships, set social standards, and negotiate and assert our society's values. In the process, they have become safe spaces for the spread of hate, conspiracy theories and disinformation.

Social media companies erode basic human rights and civil liberties by enabling the spread of online hate and disinformation.

At CCDH, we have developed a deep understanding of the online harm landscape, showing how easily hate actors and disinformation spreaders exploit the digital platforms and search engines that promote and profit from their content.

We are fighting for better online spaces that promote truth, democracy, and are safe for all. Our goal is to increase the economic and reputational costs for the platforms that facilitate the spread of hate and disinformation.

If you appreciate this report, you can donate to CCDH at counterhate.com/donate. In the United States, Center for Countering Digital Hate Inc is a 501(c)(3) charity. In the United Kingdom, Center for Countering Digital Hate Ltd is a nonprofit company limited by guarantee.

CONTENTS

1. INTRODUCTION

Social media has revolutionized an enormous array of interpersonal interactions: how we find and connect with other people; access and spread information; build relationships and community; and conduct business. It has also increased the spread of hate speech, disinformation, and threats against children and marginalized populations. These negative externalities are amplified because the platforms' commercial interests overwhelmingly take precedence over social responsibility.

The reason for this is a US law passed in 1996 – Section 230 of the Communications Decency Act – and the culture it has inculcated in Silicon Valley. Section 230 immunizes social media platforms from liability for the harms created, accelerated, and broadcasted to billions on their platforms. This unique legal immunity from liability for their products – a Get Out of Jail Free card that no other industry in America enjoys – has created a culture among social media executives in which they refuse to take any responsibility for their platforms' harms, even when they are broadcasting eating disorder content to children on a repeat loop, or seeding disinformation that leads to violence and atrocities in the real world.

The status quo is untenable. An increasing array of voices are demanding change, drawing on lived experience from myriad perspectives. Policymakers must act to stop the spread of hate speech and disinformation. Informed by our research and advocacy, we created the STAR (**S**afety by Design, **T**ransparency, **A**ccountability, **R**esponsibility) policy framework. STAR is an integrated approach to regulation where each component reinforces and enables the others to create a better online ecosystem. STAR is not reinventing the wheel for regulation—it is simply demanding that the social media industry be regulated like every other industry.

What does a STAR internet look like?

Imagine an internet where user safety and well-being are hard-wired priorities, transparency and user safety are non-negotiable, and corporate accountability and responsibility to the public come before profits.

In this STAR-compliant future, every platform, app, and online service has safety ingrained into its design and development cycles. From the start, risk assessments would incorporate user perspectives to proactively identify potential harms. Exploitative engagement tactics and dark patterns are prevented before ever reaching users.

Transparency is the norm, with platforms providing visibility into their algorithms, content moderation rules, data practices, advertising systems, and any other systems that impact user experiences. Independent regulators are empowered to hold companies accountable and responsible for their conduct.

In this STAR future, children are shielded from online harms and develop healthy digital literacy. Marginalized groups are protected from harassment. Users can make informed choices without being exposed to massive amounts of disinformation.

Under STAR, social media companies are incentivized to create safe platforms and be transparent about their product designs because they no longer enjoy complete unaccountability under Section 230. In this future, the countless stories of people experiencing online harassment and exposure to harmful content no longer occur because social media companies finally face the consequences of their actions.

Implementing STAR will usher in a digital ecosystem where child safety, human rights, civil liberties, transparency, and corporate accountability are prioritized as core considerations alongside commercial interests and profit. The time to establish these principles is now.

Imran Ahmed, CEO, Center for Countering Digital Hate

2. THE STAR FRAMEWORK

The STAR Framework is a blueprint for policymakers to combat online harms and fortify democracy in the digital age.

Safety by Design is a principled approach to the design of technology products and social media platforms that promotes user health, well-being, human rights, and civil liberties. However, safety cannot be achieved without transparency into the algorithmic systems and economic incentives driving platform features and behaviors.

Transparency is a social media company's obligation to disclose accurate and accessible information about algorithms, product design, platform decisions, and economics, particularly around advertising. Transparency efforts must be reinforced with accountability and responsibility measures like regulation, penalties for violations, and independent oversight to enforce obligations. Only then can platforms be compelled to provide the necessary visibility.

Accountability and Responsibility underscore that platforms must take ownership of their decisions and be responsive to users and democratic institutions. Governments must implement economic consequences for inaction to counterbalance the profit motives that lead companies to deprioritize safety and transparency.

Accountability and responsibility cannot be achieved in the US without reforming Section 230 of the Communications Decency Act. For too long, Section 230 has granted near-complete immunity from liability for social media companies. Meaningful safety, transparency, accountability, and responsibility are simply impossible without Section 230 reform.

Critically, accountability and responsibility cannot be fully realized without the transparency to identify issues and safety by design principles to prevent harm. STAR principles are mutually reinforcing, forming a framework to create a digital world where safety, respect for human rights, transparency, and accountability are mandatory, rather than expendable in favor of commercial interests.

This report will discuss each component of STAR, offering an in-depth analysis of the current digital ecosystem, and offer policy solutions. The last section examines how various jurisdictions have used the STAR framework to regulate the social media industry to date in August 2024.

3. SAFETY BY DESIGN

Safety by Design is a principled approach to the design of technology products and social media platforms with the primary aim of promoting user health and well-being.

A company's incentives determine how its products are designed and deployed. Social media companies are incentivized by metrics like increasing daily active users, prolonging screen time, and generating more clicks to serve as many advertisements as possible.[1] This has created a race to the bottom, where companies compete for the attention of users and use algorithms to keep people on screens longer. Social media is designed to capture as much engagement as possible, even when excessive engagement poses serious mental health and safety risks for users.[2]

Governments have stepped in to regulate industries and advocate for consumer safety and rights, be it food safety, automobile safety, or others. Today, governments must step up to regulate social media like any other industry. In this section, we discuss the existing landscape of efforts made by social media companies and governments to make platforms safer and conclude with policy recommendations for stakeholders to consider.

3a. The current landscape

Reporting on the harms posed by social media platforms has increased the public's awareness of online harms and there is clear support for safety by design.[3] As a result of scandal, public criticism, and scrutiny from governments, social media companies have begun implementing some safety features into their services. Social media companies have attempted to address safety risks by:

Table A. Platform-side safety features

Policy	Description
Community rules	Community rules are policies that set the parameters for acceptable content and clarifies categories of offensive or harmful content.
Automated review of content	Content is often reviewed by artificial intelligence (AI) for compliance with a platform's community rules.
Human review of content and appeals	Content is often further reviewed by human content moderators for compliance with a platform's community rules.
User flagging and reporting of content	Users can often flag or report content they believe may violate a platform's community rules.

Policy	Description
Safety centers and other resources	Safety centers are collections of resources users can consult to potentially receive help and understand platform rules.
Collaborative technology projects	Social media companies may collaborate to develop shared technologies that help respond to harm.

Table B. User-side safety features

Feature	Description
Age gating	Users are required to report their age when signing up and must be over a specified age to use the platform, typically thirteen.
Nudges to change conduct	Users may automatically receive notifications encouraging them to adopt healthier behaviors, such as logging off late at night.
Sensitive content controls	Users may often toggle settings to strengthen filters against content that may be upsetting or offensive.
Parental controls	A parent may be able to link their account with their child's account to monitor behavior and control account settings.[4]
Privacy settings	Users can often toggle settings that control who may contact them and view their content.[5,6]
Chronological feeds	Users may be able to halt personalized content recommendations.[7,8]

These systems and tools are not enough to mitigate online harm. In public statements, social media companies often represent these measures as being sufficient or at least contributing to addressing online harms, but internal company records paint a different picture.[9]

Platforms exaggerate the effectiveness of AI moderation.

For instance, in 2021, one of Meta's senior research scientists estimated that the company's AI tools caught content that was responsible for just 2% of all the views of hate speech on the platform, and a separate team concluded that the company's automated systems removed content that generated just 3% to 5% of views of hate speech.[10] While Meta employees were calling attention to the limits of AI content moderation, senior leadership was publicly claiming that its AI tools proactively detected 98% of all violating content.[11]

In this environment, without meaningful transparency, the public cannot validate the truth of what social media companies have to say about their safety measures, nor have any way to guarantee that they are implemented proactively and vigorously enough to address online harms.

3b. Why are the current safety measures not enough?

1. **The process of designing safety features is highly reactive and emphasizes band-aid solutions to harm.**

   Social media companies historically have appeared to rarely build their products with user health and wellbeing in mind from the beginning. Instead, social media companies tend to roll out safety features in response to public outcry or pressure from elected officials, such as hearings before a jurisdiction's legislature.[12]

   For example, in July 2023, Discord announced the creation of new parental controls that allow parents to monitor their child's account.[13] This came after explosive reports documenting that predators were using Discord to groom and abduct teens.[14] Years earlier Discord had publicly refused to develop parental controls and seemingly only changed course after the public revelations about child sexual abuse on the platform.[15] However, this safety feature is inadequate. In March 2024, Discord disclosed to the U.S. Congress that less than 1% of underage users' parents use the company's parental controls.[16]

2. **Safety features are often opt-in and place the onus on users to protect themselves from harm.**

   After safety features are rolled out, they may see little adoption because they are not turned on by default, have not been actively promoted to users, or are difficult to activate.

   For instance, in 2019, Instagram promised to implement "Project Daisy", a version of the app in which Likes were turned off by default.[17] According to discovery findings from Meta's litigation with U.S. state Attorneys General, internal research found that turning off "Likes" led to significant benefits for users' mental health and well-being and that the feature was projected to lead to a 1% decrease in advertising revenue.[18] Unfortunately[19] Turning off "Likes" was rolled out as a feature that users could opt into, and only accessible after navigating.[20]

3. **Safety measures are poorly maintained and implemented.**

   Social media companies have approached safety measures as one-off, check-the-box projects instead of dynamic, ongoing initiatives that see substantial investment and improvement over time.[21] As a result, once a project is completed, it may receive less attention and maintenance than other types of projects. This means safety measures may be prone to bugs and errors.

For instance, over a period of six months in 2020 and 2021, Meta failed to take down millions of child abuse images due to two technical errors in the system the company uses to detect child abuse content.[22] In 2019, a similar error occurred that also resulted in millions of child abuse images going undetected.[23] Though the bugs were eventually discovered, it took Meta months to do so, and only after the lower detection rate had been disclosed in its quarterly transparency reports.[24]

In another example, Meta rolled back protective measures after the 2020 election designed to counter misinformation, leaving the company unprepared for the events of January 6th, 2021.[25]

3c. Policy solutions

Safety by design should be adopted for all social media platforms. Regulations are needed to incentivize social media companies to design their products with user health, well-being and human and civil rights integrated from the beginning and to continually review their products for safety risks. In addition, users should be given greater control over their personal data and content recommendations.

Policymakers have the power to mandate safety by design. This section lays out a suite of options for governments to consider, organized into four categories: reorient the product design process; empower users; create robust systems for tackling online harms; and protect minors online.

Reorient the Product Design Process

- **Enshrine duties of care to address online harms.** A "duty of care" would require tech companies to exercise reasonable care in the design of their services to avoid foreseeable harm to users, particularly children.
- **Set standards for safe product design.** Industry standards create a clear baseline of expectations for product design. For example, standards should require services to enable their strongest privacy settings and content filters by default.
- **Mandate risk assessments and mitigation plans.** Risk assessments involve systemic evaluation of the effects social media platforms have on users' health and wellbeing. Risk mitigation plans should clearly enumerate steps to address the identified risks.
- **Independent audits by third parties and regulators.** Risk audits enable an added layer of scrutiny of a digital service's design. For audits to be effective, trusted third parties must have enough information to understand the inner workings of a product.

Empower Users

- **Restrict the usage of manipulative design features, including deceptive engagement patterns, also known as 'dark patterns'.** Dark patterns are design features that disempower users and manipulate their online experience. This includes but is not limited to inducing users into staying online longer through autoplay, overloading users with multiple requests, forcing users to navigate through multiple pages to access policies or information. These design features should be restricted.
- **Incentivize features that encourage healthier forms of engagement.** Social media platforms can nudge users into taking breaks and reconsidering abusive posts. These features should become an integral part of the user experience.
- **Allow users to permanently delete their accounts and data.** Users should be able to opt out of social media platforms at any time. When users feel unsafe or unhealthy online, they should be able to opt-out by deleting accounts and data.

Create Robust Systems for Tackling Online Harms

- **Invest significant resources into trust and safety.** Tech companies must contribute the necessary funding and staffing to mitigate harmful content and to proactively detect harmful, violative, and illegal content.
- **Establish specialized reporting pathways for the most egregious harms.** Companies should build specialized processes to address the worst harms, for instance, a pathway to report unwanted sexual advances or flagging child sexual exploitation and abuse. Platforms should also establish pathways for victims of targeted harassment campaigns.
- **Create protocols and systems for engaging with law enforcement.** Law enforcement should be able to notify tech companies of illegal content through an accessible system, with the ability to review notices and clarify actions taken.

Protect Minors Online

- **Examine and study safety risks to minors.** Duties of care and risk assessment of product design and algorithmic recommendations that pay special attention to the needs of minors who evidence has shown are susceptible to different online risks than adults.
- **Create safeguards and robust parental controls.** While manipulative design features should be restricted for all users, this is doubly true for minors. Parents should be empowered with tools to ensure their children's safety online.
- **Establish a dedicated reporting mechanism for minors.** When minors report online harms, they should be given priority review and response by tech companies. This can be accomplished through establishing specialized reporting pathways for minors.
- **Protect children's personal data and restrict targeted ads.** When minors use social media platforms, their personal data should receive the strongest protections and never be processed for targeting ads.

4. TRANSPARENCY

Transparency is a digital service company's obligation to disclose accurate and accessible information about algorithms, product design, platform decisions, and economics, particularly around advertising.

Transparency is a central pillar of comprehensive internet policy reform. Transparency is how we can make the platforms responsible for their negligence and hold them accountable to the public. Currently, the platforms have no incentive to define meaningful transparency—just look at how they continuously withdraw from voluntary transparency tools.[26] Policymakers need to step in and create a transparency framework that delineates who should receive information, what information should be shared, and how it should be shared.

This section discusses the existing landscape of transparency and information-sharing practices by companies and legislative efforts by governments and concludes with policy recommendations.

4a. The current landscape

The platforms provide minimal transparency through two main methods: user-facing transparency and technical transparency, outlined below.

Table C. User-facing transparency

Practice	Description
Community rules	Guidelines or policies that dictate how users should behave on a platform, outlining processes for reviewing content and appealing decisions.
Terms and conditions agreements	Upon sign up, users typically visit a document that contains the rules of use for a website and certain disclosures, such as how personal data is processed.
Transparency centers	Social media companies often publish quarterly or biannual reports detailing community standards enforcement, government data requests, or intellectual property violation. They also on occasion publish internal research on product testing.[27]
Advertising standards and disclosures	Guidance on the type of advertisements allowed or prohibited on the platform and clear disclosures of advertisements on the platform.

Table D. Technical Transparency

Practice	Description
Data access tools	Interested parties may be able to request specific tranches of granular information using dedicated tools, usually APIs
Ad and content libraries	The public may use databases to search for information about advertisements and user-generated content.
Collaborations with academics and researchers	Platforms may partner with select academics to study aspects of their services, such as their effect on voting behavior.

Current transparency measures are misleading.

While these information-sharing practices may seem comprehensive, in practice they are often nothing more than an exercise in public relations for social media companies. As long as the companies have complete control over what data they make public, these transparency measures are designed to protect the interests of companies.

Over the past years, whistleblowers have come forward and disclosed shocking revelations about the conduct of social media companies.[28]

In 2023, Arturo Bejár, Meta's former head of online safety, testified before Congress on the findings of internal research he conducted in 2021 surveying users about their experiences with online harms on Instagram.[29] The results were stark:

- **30.3%** of users said they saw misinformation.
- **25.3%** witnessed hate.
- **11.9%** received unwanted sexual advances.
- **6.7%** were exposed to self-harm content.[30]

Internally, Meta's leadership chose to suppress Bejár's findings, requiring him to write about safety risks as if they were "hypothetical".[31] At the same time, Meta was publicly claiming in its transparency reports that the prevalence of hate speech was just "0.05%" and that it had "decreased for three quarters in a row".[32]

This discrepancy between internal research and Meta's representations of harmful content is exactly why robust transparency is critical for regulating digital service platforms.

4b. Why are the current transparency measures are not enough?

1. The myth of "prevalence"

Social media companies selectively share information without full context to appear transparent. For instance, some social media companies report statistics about the removal and spread of harmful content using a metric commonly known as "prevalence". Prevalence is calculated by estimating how many views were received by content that violated community rules and then dividing by the total number of views on a platform.[34]

Though superficially reasonable, prevalence is designed to let social media companies deny the problem and delay action. It masks high rates of harmful content behind a huge denominator (the total views on a service) and fails to include content not explicitly barred by a company's community rules, which may be narrowly defined. It also communicates little about specific types of harmful content and the communities that are impacted by it.

A more useful and accurate way of assessing the rate of harmful and violative content is through 'discoverability.' Discoverability is the ratio of violative content to the relevant sensitive content.[35] For example, to assess the rate of health misinformation, it would be more descriptive to divide the views of health misinformation over the total number of views on the specific health topic instead of the total views across the whole platform.

2. Data access tools have been closed, limited to select groups, and designed to include frictions that inhibit usability.

Data and technical information about content and a product's design are some of the most valuable resources for understanding online harms. In recent years, social media companies have restricted, eliminated, or changed the tools researchers use to access data.[36] In 2023, Twitter (now X) systematically cut off researchers' ability to access data via its API, previously a common source of social media data for research.[37] The price for the previously free service was raised to $42,000 per month, disrupting hundreds of independent research projects that relied on Twitter data. X's new leadership has even gone on to sue several independent researchers, including CCDH. A survey of independent researchers found that a majority of respondents fear being sued by X over their findings or use of data.[38] X is not the only platform to engage in this behavior: in 2021, Meta abruptly cut off researcher's access to transparency tools.[39] Earlier this year, Meta announced that it would shut down CrowdTangle, a platform monitoring tool that was used by independent researchers.[40]

3. Transparency efforts are inconsistent, subject to change, and opaque.

Social media platforms and the internet have come to define how we communicate in the modern world, becoming the "modern public square."[41] Yet these modern public squares set rules without democratic input, arbitrarily enforce them, and then refuse to provide meaningful transparency to the public.

Transparency is valuable when it is both lasting and standardized, allowing for comprehensive research into the social media industry. Many platforms publish community rules and standard procedures for reviewing content, but they are not fully transparent about how they enforce their rules.

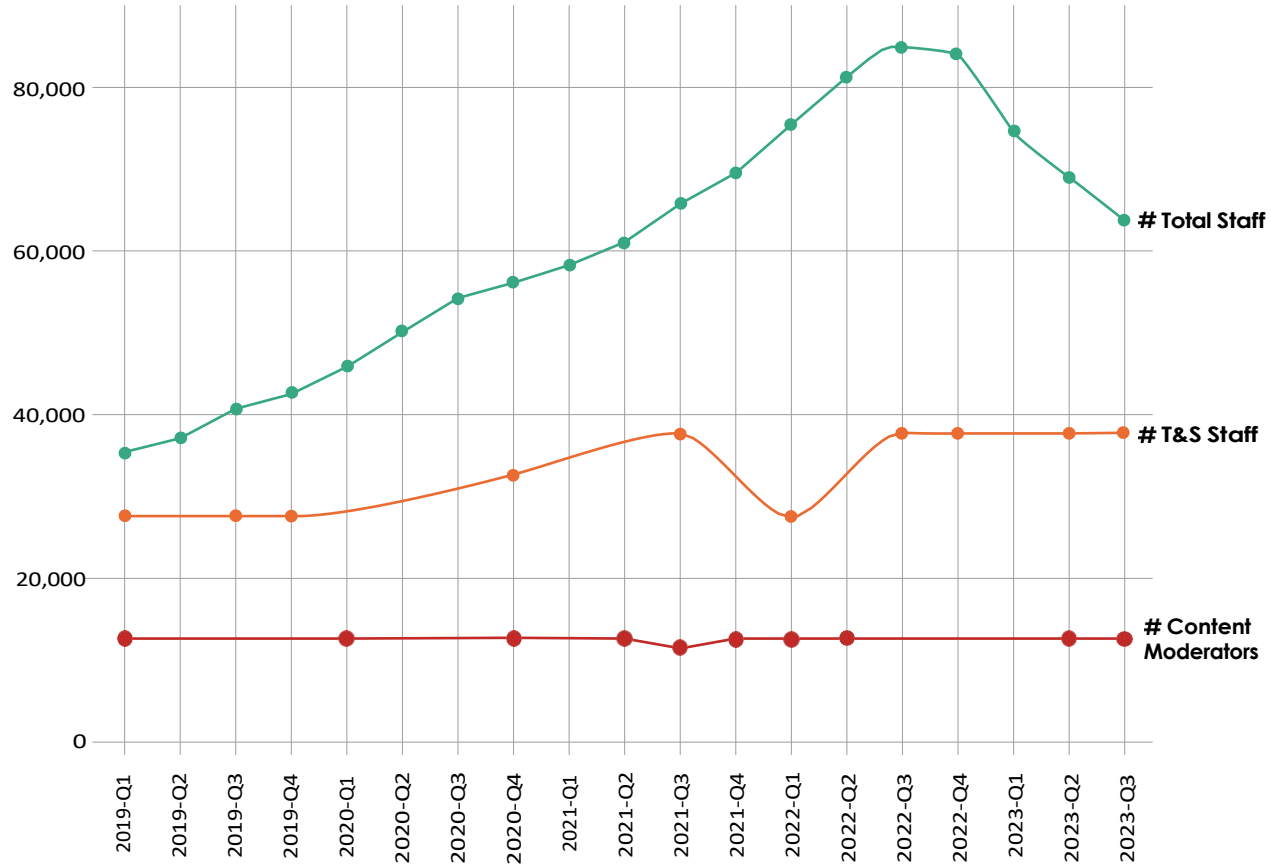
With no sector-wide standards, platforms choose their own approach resulting in a variety of voluntary transparency reports that provide varying levels of useful information. For a long time, Snap (the owner of Snapchat) restricted access to its content guidelines to vetted publishing partners and professional creators. It was not until May 2023 that Snap made its content guidelines public.[42] This stands in stark contrast to other social media companies, such as Meta and YouTube, which have disclosed their community rules for years.[43]

4. External audiences have few pathways to verify the accuracy of disclosures or request additional information.

Voluntary information-sharing practices are meaningless for external audiences because social media companies have little incentive to be transparent about their products' internal workings.

An illustrative example of this problem is social media companies' practice of outsourcing content moderation.[44] In recent years, social media companies have disclosed that they hire tens of thousands of content moderators and trust and safety specialists.[45] However, basic facts like the number of content moderators they hire remain unclear. And thus, it is unclear how much platforms invest in trust and safety. For instance, Meta has claimed since at least 2019 that it employs around 15,000 content moderators worldwide despite considerable fluctuation in its overall headcount.[46]

Meta Staffing Levels 2019-2023



4c. Policy solutions

Transparency is an essential component of any online safety regime. A regulatory body that can lead the creation and development of a transparency framework is critical for sustainable transparency, instead of creating one-off measures.

This section gives an array of options for policymakers to consider, broken down by the target audience of transparency: the public, regulators, and researchers.

Transparency for the public

- **Require the creation of "acceptable use policies."** These policies should clearly disclose the types of content that is prohibited, how content is reviewed, how users' data is utilized by the platform, and the steps users must take to report potentially violating content and appeal decisions made by the platforms.
- **Mandate the publication of standardized regular transparency reports.** Transparency reports should contain statistics, native analytics, and top-line summaries about the quantity of content flagged by users, actions taken, user appeals, engagement with content, trust and safety spend, and other relevant data.

- **Publish comprehensive content libraries.** Content libraries should contain robust information about viral content, monetization policies, content created by major public accounts, and paid advertisements. Such data should be made easily searchable and downloadable.
- **Create robust legal protections for whistleblowers.** Former employees of social media companies who disclose useful information about harmful corporate conduct should be protected from retribution.

Transparency for regulators

- **Establish an independent authority to enforce transparency.** Lawmakers should create a well-resourced, expert unit of government empowered to set standards for transparency, ensure access to data, and authority to request and compel information from companies.
- **Require disclosure of emerging trends, mitigation plans, and independent audits.** To ensure compliance, social media platforms should be required to share documentation of their efforts to promote user health, well-being, human and civil rights with regulators and the public.
- **Oblige companies to disclose information about product design.** Upon request, social media companies should disclose information about their algorithmic design, product design, native analytics, and optimization metrics.

Transparency for researchers

- **Establish a research program to study online harms.** The independent authority should collaborate with civil society and academia to study online harms in the public interest. It should be charged with approving projects and ensuring access to all necessary data.
- **Issue standards for data access.** The independent authority should set standards for how platforms make data and all relevant documentation available, including via APIs.
- **Certify researchers and ensure robust protections for data privacy.** The independent authority should establish a process for training and vetting researchers in the safe use of social media data and issue mandatory guidelines for protecting data privacy and security.
- **Enshrine strong liability protections for independent researchers.** Research studying online harms in the public interest should not be chilled by litigation. Researchers should be protected at the earliest stage of the legal process except when data is misused.

5. ACCOUNTABILITY, RESPONSIBILITY, AND CRITICALLY, SECTION 230.

Accountability is the obligation of social media companies to explain and justify their safety measures and company practices to public institutions, and responsibility is their duty to protect and be responsive to users.

Social media companies can and should design their services with user safety in mind. To be accountable to institutions and responsible to users, laws and regulations must impose consequences on social media companies for failing to uphold their duties to create safe products.

Social media companies have been able to evade accountability and responsibility due to Section 230 of the Communications Decency Act, an outdated U.S law that shields social media companies from liability. Because the social media industry is primarily based in the United States, American laws that govern the industry have an international impact. The global dominance of American social media companies and the lack of regulations has exported social media platforms that are rife with hate speech, misinformation, and disinformation.

When online safety laws specify how social media companies should be transparent and safely design their services, public institutions can ensure these expectations are met by monitoring, requesting information about, and penalizing illicit corporate conduct.

This section explores how social media companies have exploited the absence of accountability and responsibility, how some jurisdictions have established institutions to oversee corporate conduct and concludes with recommendations for policymakers to consider.

5a. The weaponization of Section 230

Following the emergence of websites that allowed users to directly engage in forums, the 1996 Communications Decency Act became law in the U.S.[47] A key section of the law commonly known as Section 230 immunized "interactive computer services" from liability for damages caused by their product. Section 230 has two important sub-sections, each handing social media platforms powerful tools to avoid accountability.

The first sub-section, (c)(1), immunizes platforms from lawsuits that treat them as the "publisher or speaker" of content created by 'third-party' users. This means that, for instance, the parents of children who overdose on drugs obtained via social media apps cannot sue the platforms because Section 230 holds that platforms have no responsibility for content created by drug dealers on their apps, even if their algorithms do nothing or promote it in children's feeds.[48]

The second sub-section, (c)(2), shields companies from liability for decisions they take on what they allow users to share on their platforms. This sub-section is known as the "Good Samaritan" portion of Section 230 because its intent was to encourage websites to proactively remove

harmful content. However, while companies can and do use this freedom to proactively remove objectionable content, nothing obliges them to do so. This means, for instance, social media companies may or may not choose to remove deepfake pornography of women on their sites without the risk of facing lawsuits from its creators.[49]

The internet was vastly different in 1996 and the law was meant to protect a burgeoning industry from excessive litigation. However, today the social media companies are a multi-trillion-dollar industry that use Section 230 to evade liability and externalize harms caused by their products. Table E. lists some of the cases of how social media companies use Section 230 to deny victims redress for online harms:

Table E. Section 230 Enables Companies to Silence Victims and Avoid Responsibility

Digital media company	How It Weaponized Section 230
Grindr	Grindr used Section 230 to dismiss a lawsuit by Matthew Herrick, who was stalked and impersonated on the dating app.[50] Herrick sent Grindr over one hundred complaints, but the app refused to act.[51]
TikTok	TikTok used Section 230 to dismiss a lawsuit by a mother whose daughter died after attempting a dangerous "blackout challenge", which was suggested by TikTok's algorithm.[52]
Snapchat & Yolo	Snapchat and Yolo used Section 230 to dismiss a lawsuit by Kristin Bride, whose son Carson died by suicide after receiving hundreds of abusive messages on the anonymous Yolo app.[53]
Facebook & Instagram	Meta is currently trying to use Section 230 to dismiss a lawsuit by parents, schools, and 33 Attorneys General, who argue that Instagram and Facebook's design is addictive and harmful to children.[54]
YouTube & Reddit	YouTube and Reddit tried using Section 230 to dismiss a lawsuit by Black victims of a mass shooting in Buffalo, New York, who argued that YouTube and Reddit's algorithms are a "defective product".[55]
Amazon	Amazon used Section 230 to dismiss a lawsuit by parents whose children bought "suicide kits" (concentrated sodium nitrite) on its store, which they used to commit suicide.[56]
X	X used Section 230 to dismiss a lawsuit by two men, who had been sexually trafficked as minors and whose abuse materials had been allowed to spread on the platform.[57]

5b. How international governments have provided reasonable liability protections while balancing consumer safety.

The EU's e-commerce directive, enacted in 2000, set the foundations of the EU's liability regime for 'intermediary service providers. Inspired by Section 230, it exempted liability for services providing hosting, caching and 'mere conduits' and covering all illegal content and activity. Unlike its US equivalent, the exemption is conditional on a 'reasonableness standard,' i.e., the removal, or disabling of access to content once companies become aware of its existence. This limit to the liability exemption allowed the EU to later introduce stricter rules for companies without changing the underlying principle.

It is in this context that the EU's 2023 Digital Services Act attempts to both strengthen the legal certainty offered by liability exemptions and increase accountability of companies. On one hand, it makes clear that platforms are under no obligation to proactively search for illegality and are exempt from liability if they undertake voluntary measures 'in good faith and in a diligent manner' (the Good Samaritan principle). On the other hand, it attempts to limit the possibility for a platform to remain unaware of illegal content by requiring robust user reporting and the introduction of 'trusted flaggers'. Backed by the ability to levy large fines and restrict market access, the EU also demands an extensive duty of care from companies. This includes demanding transparency, researcher access, risk assessments and the mitigation of systemic risks which encourage safety by design.

Having entered into force in early 2024 and with early enforcement actions showing promise, the DSA is poised to pressure social media companies into prioritizing user well-being by putting a cost on irresponsibility. This 'accountability not liability' approach demonstrates an alternative path to regulating social media without fundamentally altering the underlying legal framework.

5c. The current landscape and the myth of self-regulation.

Social media companies have occasionally created mechanisms to hold themselves accountable and responsible. These mechanisms are advisory councils or external organizations, such as X's now-defunct Trust and Safety Council and Meta's Oversight Board.

Companies claim that these bodies are meant to help them receive feedback from external audiences or independent checks on their content moderation decisions, but in practice they assist more with public relations than with substantive improvements in trust and safety.[58]

For instance, X's Trust and Safety Council originally included around one hundred civil society organizations and experts on online harms.[59] However, following Elon Musk's acquisition of the company in 2022, it was disbanded less than an hour before its final meeting, and no replacement has been announced since.[60]

In 2020, Meta announced its intention to create and fund an "Oversight Board" – akin to a "Supreme Court for Facebook" – comprised of experts and representatives of civil society to review and give recommendations for improving its content moderation.[61] Meta committed to abide by the Board's determinations and guarantee its independence, extending it $280 million in funding via an irrevocable trust. The Board's design and decision-making process have critical deficits, making it an inadequate substitute for meaningful government regulation.[62] The Board's authority to review Meta's content moderation is weak and limited. It can only review moderation of individual posts, and its recommendations are nonbinding, hindering its ability to genuinely influence broader rules about content.[63]

Meta and the Board take months to go back and forth about pressing, even dangerous pieces of content, such as a video featuring Cambodia's authoritarian former Prime Minister Hun Sen inciting violence against his political opponents.[64] It took Meta over six months to remove the video. In addition, Meta claims to listen faithfully to the Board's recommendations, but many that are listed as being implemented on its website remain "in progress" years after their issuance.[65]

The Board can review any of the millions of Facebook and Instagram posts moderated every day, yet it has issued a paltry number of decisions since its establishment.

And while the Board's judgments may reverse decisions by Meta, they seem calculated to generate headlines. It has actively dodged reaching decisive determinations in the most controversial cases while taking pains to seem to contradict Meta. This behavior helps the Board maintains a veneer of legitimacy as an oversight body, while Meta appears compliant with an external source of authority.[66]

5d. Policy solutions

Accountability and responsibility are essential elements of any law intended to change the behavior of social media companies. When companies are fully shielded from liability, they have little reason to address safety risks or disclose potentially embarrassing information to the public. Governments must establish and empower public institutions with a mandate to impose consequences on social media companies. To that end, public institutions must be equipped with the necessary legal authorities, resources, and expertise to hold them accountable. These are key components of any effective, dissuasive regulatory framework.

- **Reform Section 230 following the 'deliberate indifference' standard.**
  Legal scholar Mary Anne Franks has suggested that Section 230 (c)(1) immunity for intermediaries should only be granted when three conditions are met: one, when the content in question is speech, as opposed to conduct; two, when the speech is wholly provided by a third party, as opposed to being solicited or encouraged by the platform itself; and three, when the platform has not exhibited deliberate indifference to harm caused by that speech.67 This reform would narrow the scope of online activity for which platforms could avoid accountability, thus incentivizing them to act more responsibly, without exposing them to strict liability for the speech of others.

Here is how we would reform Section 230 to balance reasonable liability with consumer safety:

1) **Treatment of publisher or speaker**

No provider or user of an interactive computer service shall be treated as the publisher or speaker of any ~~information~~ speech wholly provided by another information content provider, unless such provider or user demonstrates deliberate indifference to harm caused by that speech.

2) **Civil liability**

No provider or user of an interactive computer service shall be held liable on account of-

A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is

constitutionally protected; or

B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1);1

3) **Limitations**

The protections of this section shall not be available to a provider or user who manifests deliberate indifference to unlawful material or conduct.

- **Condition protections from liability on minimum standards of behavior.** Blanket immunity creates little incentive to proactively address online safety risks. Governments should condition immunity on the expectation that social media companies take reasonable steps to do so.
- **Establish an independent digital regulator dedicated to online safety.** Independent regulators are the best actors to enforce and update online safety requirements. This model permits both flexibility and the development of expertise as technology changes over time.
- **Give regulators strong investigative authorities.** Regulators should be able to request and demand information from platforms about online harms. Companies should be required to respond to requests within a fixed period. This ability is essential for effective enforcement.
- **Enable regulators to levy fines and fees.** Regulators should be enabled to issue dissuasive fines for noncompliance with legal requirements and fees to fund enforcement. This ensures that illicit behavior hits the bottom line and companies bear the burden of enforcement costs.

6. STATE OF STAR

Social media is global and CCDH operates in several jurisdictions to advocate for comprehensive social media reform. This section assesses how governments globally have applied the STAR framework to regulating social media companies as of August 2024.

6a. The United States

Safety by Design

Since the internet's inception, U.S. federal policy has refrained from promoting safety by design, though interest among federal legislators and state policymakers has climbed in recent years. Safety by design efforts in the U.S have been aimed at protecting children, not the overall public.

In 1998, the Children's Online Privacy Protection Act (COPPA) became law, making it the only piece of federal regulation promoting safety by design in the last 30 years.[68]

Recently, legislative proposals promoting safety by design for children, such as COPPA 2.0 and the Kids Online Safety Act (KOSA) have received widespread strong bipartisan support.[69] In the absence of strong federal regulations, state legislatures have been tackling safety by design, passing laws that address user data and promote child safety.[70]

Transparency

Unlike other jurisdictions, the US has not passed a sectoral law requiring social media companies to be more transparent, though regulators have used existing authorities to scrutinize social media companies' public claims. Recently, lawmakers have introduced bi-partisan legislation aimed at strengthening transparency requirements for social media companies.[71]

In the US, consumer protection law bars companies from engaging in "unfair" or "deceptive" business practices and empowers a generalist agency, the Federal Trade Commission (FTC), to investigate and launch enforcement actions, including against social media and other tech companies.[72] For instance, in 2019, the FTC levied a $5 billion fine against Meta, the largest in the agency's history, for deceiving consumers in the wake of the Cambridge Analytica scandal.[73]

Accountability and Responsibility

In the US, policymakers have taken a hands-off approach to accountability and responsibility, trusting companies to follow through on their commitments to address online harms and self-correct when gaps become known. Congress has used public hearings to hold social media companies CEOs "accountable" in absence of real regulation.[74]

Section 230 epitomizes this approach. The law was intended to enable well-intentioned companies to remove objectionable content without fear of legal repercussions from its creators or victims. This "Good Samaritan" protection assumed that companies were trustworthy actors that would willingly take responsibility for online harms.

Policymakers have also entrusted an existing regulator – the FTC – to use its authorities under consumer protection law to hold social media companies to their public promises. When a company breaks one of its promises, the FTC can impose consequences like fines and warning letters, but it is questionable whether this framework is sufficient to deter such behavior.

6b. The United Kingdom

Safety by Design

The UK's Online Safety Act (OSA) became law in 2023. The OSA places legal responsibility on social media companies to deal with illegal content, like terrorism and revenge pornography, and to stop children seeing harmful material such as self-harm, eating disorder promotion, and pornography.[75]

Throughout the legislative process, witnesses and parliamentarians referenced Safety by Design as the principal aim of the Online Safety Act's requirements.[76] The insistence on safety by design principles led parliament to focus the Act on addressing systems and processes rather than identification and takedown of violative content. Systems and processes regulation is more proactive than requirements like identification and takedown because it addresses root causes and platform design rather than individual pieces of content.

The Online Safety Act is in the process of being implemented by the UK's independent online safety regulator, Ofcom. It is vital that Ofcom upholds the will of Parliament, which was clear in identifying safety by design as the guiding principle of the Online Safety Act.

Transparency

The OSA places transparency and reporting requirements on regulated social media platforms and search services. Many of the legal duties of care the Act establishes have been linked to transparency reporting requirements in which platforms must include information about their adherence to the Act's duties in a form accessible and assessable by the regulator. All services have a duty to keep written records, in easily understandable form, of the risk assessments they undertake and the measures they implement to comply with their duties under the regulator's code of practice recommendations.

Ofcom has broad information gathering powers to force transparency from social media platforms and search services under its purview. However, transparency and data access powers for researchers and other independent overseers are lacking in comparison to other

jurisdictions, such as the European Union's Digital Services Act. A late amendment to the Online Safety Bill in July 2023 stipulated that Ofcom must undertake a review of research access to platform data and author a report on the transparency of these companies to independent oversight within 18-months of the regulatory regime coming into force.[77]

Accountability and Responsibility

Through the OSA, the UK has taken important steps towards ensuring that platforms have increased legal responsibility for the safety of their users and accountability to public institutions, in this case the regulator Ofcom. The OSA places a duty of care on platforms, requiring they protect all users from illegal content and child users from additional categories of content that is harmful to them. Ofcom is empowered to oversee and enforce these duties, judging whether the measures platforms take meet the duties required in law.

Should the regulator judge a platform to be in non-compliance with its legal responsibilities, the Act grants broad powers to investigate, request information from and inspect online service providers. Should these inquiries fail to produce compliance, or even greater failings are discovered, Ofcom can impose fines of up to £18 million or 10% of worldwide annual revenue, whichever is greater.

Further, the Online Safety Act also created new criminal offences, including liability on senior managers at regulated platforms should they fail to comply with the regulations. Under Section 110, a named individual acting as a senior manager can be prosecuted for failing to comply with information request from the regulator or for failing to take reasonable steps to comply with the Act.[78]

6c. The European Union

Safety by Design

The European Union's Digital Services Act (DSA)[79], adopted in 2022 and entering into force in 2024, provides a harmonized approach to content moderation on social media across the 27 Member States. It imposes obligations on all intermediaries providing their services to users in the EU, with specific additional requirements for large platforms with over forty-five million active monthly EU users.

The DSA advances safety by design, focusing on prohibiting harmful practices and requiring due diligence. It encourages a broad duty of care and creates checks and balances in the product design process. Large platforms are required to identify and mitigate systemic risks stemming from their service design or functionality, clamping down on issues such as 'dark patterns' and addictive design elements.[80]

To empower users, the DSA mandates that they be given choices in how content and advertising are recommended by the platform. Additionally, users gain the right to compensation for damages or losses suffered due to DSA infringements.

Systemic risks are a focal point for the DSA and must be reflected in a platform's terms and conditions. These risks encompass illegal content, fundamental rights, effects on democratic processes, and issues stemming from the design or use of a platform. Platforms are required to conduct risk assessments to identify potential harm in these areas and ensure appropriate mitigation measures are in place.

Mitigation measures can range from general actions, such as design choices and algorithm adaptations, to more targeted interventions like age verification to protect minors and disclaimers for misleading AI-generated imagery to prevent misinformation. The DSA's emphasis on safety by design seeks to create a safer digital environment by addressing both the root causes and manifestations of harmful practices on large platforms.

Transparency

The DSA requires that large platforms publish public reports on user statistics, advertising practices, human content moderation, risk assessments and accompanying mitigation measures.[81] They are also obliged to undergo annual independent audits, the results of which must also be reported publicly.

It also addresses the information asymmetry typically hindering regulators by empowering the European Commission and national authorities with access to algorithms and any data necessary to assess the risk and harm produced by platforms. A similar asymmetry lies in the mismatch of technical expertise between the regulator and the regulated, but some progress has been made with the launch of the 'European Center for Algorithmic Transparency' which will assist with enforcement.[82] Adding further to external scrutiny, researchers are to be granted access to specific platform data upon request and when granted 'vetted researched' status by a national Digital Service Coordinator.[83]

Accountability and Responsibility

Under the DSA, the liability of platforms for illegal content is contingent on their role and the nature of their services. For instance, platforms that act merely as intermediaries—such as those offering mere conduit, caching, or hosting services—are exempt from liability for the illegal content uploaded by users, provided they meet certain conditions. These conditions include the requirement to act expeditiously to remove or disable access to illegal content upon obtaining knowledge of its existence.

The DSA's due diligence requirements provide many opportunities for enforcement by the European Commission and Digital Services Coordinators. There are also numerous opportunities in the DSA for the development of guidelines and codes of practice for regulators to respond to evolving needs and circumstances.

Failure to comply with the DSA can result in significant fines of up to 6% of global turnover, and the European Commission has already opened dozens of investigations into the world's biggest social media companies. It has to date decided to move forward with formal proceedings against X, Meta and TikTok for suspected infringements such as around the removal of illegal content, harmful design practices and data access for researchers.[84] [85] [86]

6d. Australia

The Online Safety Act 2021 (OSA) expands Australia's protections against online harm and imposes significant implications for online service providers. The law empowered an existing independent agency, the eSafety Commissioner, with the authority to regulate social media companies.[87]

Safety by Design

The OSA set out clear safety expectations from social media companies by establishing a set of Basic Online Safety Expectations and requiring the industry to develop new codes for illegal and restricted content. The OSA created enforcement guidelines on adult cyber abuse, child cyber-bullying, image-based abuse, and other types of illegal or harmful content.

Transparency

Under the OSA's guidelines, eSafety can issue periodic and non-periodic reporting notices requiring online service providers to report on their compliance with the Commission's regulations. The commission publishes the responses from the companies in transparency reports on its website for public consumption.[88]

Accountability and Responsibility

The commission is empowered to enforce civil penalties and injunctions against social media companies that fail to comply with the regulations or fail to deal with complaints in line with the platform's community policies.

7. CONCLUSION

Social media has connected the world, but it has also been weaponized to spread hate, amplify disinformation, and exploit vulnerable populations at an unprecedented scale.

STAR is a comprehensive policy framework that can realign social media—but this transition will not be simple. The lack of responsible social media governance in the United States has impacted the global behavior of digital platforms. This has created a culture of impunity among social media companies, further entrenching corporate interests.

Despite these challenges, initial steps are being made. The EU, UK and Australia have surpassed the United States in passing legislation which aligns with STAR's recommendations and principles. Now, the shift towards an online world that prioritizes people over corporations hinges on the effective enforcement of these regulations and on other governments following suit. The STAR framework sets out a path for policymakers to demand the same level of accountability from social media companies as we expect from other industries. Its implementation would represent a revolution in reasonableness putting an end to a status quo which poses existential risks to human rights, social cohesion, and democracy itself.

8. REFERENCES

1 "Social media platforms generate billions in annual ad revenue from U.S. youth." Harvard Chan School of Public Health. December 27, 2023. https://www.hsph.harvard.edu/news/press-releases/social-media-platforms-generate-billions-in-annual-ad-revenue-from-u-s-youth/

2 "Protecting Youth Mental Health: The U.S. Surgeon General's Advisory". US Department of Health and Human Services. December 06, 2021. https://www.hhs.gov/sites/default/files/surgeon-general-youth-mental-health-advisory.pdf

3 "Public Support for Social Media Reform". Center for Countering Digital Hate. August 2023. https://counterhate.com/wp-content/uploads/2023/08/STAR-Report_FINAL.pdf

4 "Parental Guide for Teens on Instagram". Instagram. Accessed June 12, 2024. https://about.instagram.com/community/parents

5 "Account Privacy Settings". TikTok. Accessed June 1, 2024. https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/making-your-account-public-or-private.

6 "Instagram Help Center". Instagram. Accessed June 07, 2024. https://help.instagram.com/448523408565555?helpref=faq_content

7 "Instagram is adding a chronological feed for Reels and Stories in Europe". The Verge. August 22, 2023. https://www.theverge.com/2023/8/22/23841173/instagram-facebook-meta-chronological-feed-stories-reels-european-union-digital-services-act

8 "Meta's Threads just got an update that users have been begging for". CNBC. July 25, 2023. https://www.cnbc.com/2023/07/25/meta-threads-gets-chronological-feed-for-people-you-follow.html

9 "How Facebook uses super-efficient AI models to detect hate speech" Meta. Accessed July 4, 2024, https://ai.meta.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/

10 " Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts." Wall Street Journal, October 17, 2021, https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184

11 Ibid.

12 " Six takeaways from a contentious online child safety hearing". New York Times. 31 January 2024 https://www.nytimes.com/2024/01/31/technology/tech-senate-hearing-child-safety.html

13 "Stay connected with your teen using Discord's family center. Discord Blog. July 11, 2023. https://discord.com/blog/discord-family-center-stay-connected-with-your-teen

14 "Child predators are using Discord, a popular app among teens, for sextortion and abductions". NBC News. June 21, 2023. https://www.nbcnews.com/tech/social-media/discord-child-safety-social-platform-challenges-rcna89769

15 "Discord Chat App Is Safer Now for Kids but Still Lacks Parental Controls" The Wall Street Journal. January 19, 2021. https://www.wsj.com/articles/discord-chat-app-is-safer-now-for-kids-but-still-lacks-parental-controls-11610805602?curator=TechREDEF

16 " Fewer than 1% of parents use social media tools to monitor their children's accounts, tech companies say" NBC News. March 29, 2024, https://www.nbcnews.com/tech/social-media/fewer-1-parents-use-social-media-tools-monitor-childrens-accounts-tech-rcna145592

17 " This Is the Guy Who's Taking Away the Likes". The New York Times. January 17, 2020. https://www.nytimes.com/2020/01/17/business/instagram-likes.html

18 Unredacted federal complaint filed by 33 attorney generals against Meta Platforms, Inc. Page 49. https://oag.ca.gov/system/files/attachments/press-docs/Less-redacted%20complaint%20-%20released.pdf

19 "Behind Instagram Head Adam Mosseri's Mixed Record on Youth Safety. The Information. December 20, 2023. https://www.theinformation.com/articles/behind-instagram-head-adam-mosseris-mixed-record-on-youth-safety

20 "Instagram Help Center". Instagram. Accessed July 1, 2024. https://help.instagram.com/113355287252104

21 "Case Study on Online Youth Harms – Project Daisy". Harvard Shorenstein Center. November 23, 2023, https://shorensteincenter.org/case-study-online-youth-harms-project-daisy/

22 "Facebook blames glitch after huge drop in child abuse image takedowns". The Telegraph, May 19, 2021. https://www.telegraph.co.uk/technology/2021/05/19/facebook-blames-glitch-huge-drop-child-abuse-image-takedowns/

23 "Facebook failed to block child abuse videos after system glitch". November 12, 209. https://www.telegraph.co.uk/technology/2019/11/12/facebook-failed-block-child-abuse-videos-system-glitch/

24 "Transparency Center: Transparency Reports". Meta. Accessed June 12, 2024. https://transparency.meta.com/reports/

25 "Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs". The Washington Post. October 22, 2021. https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

26 Microsoft, Meta, Apple and Alphabet transparency tools are a 'disappointment' ahead of election, new study shows. CNBC. April 16, 2024. https://www.cnbc.com/2024/04/16/ad-transparency-tools-major-disappointment-ahead-of-election-study.html

27 "What Our Research Really Says About Teen Well-Being and Instagram". Meta. September 26, 2021.https://about.fb.com/news/2021/09/research-teen-well-being-and-instagram/

28 " The Facebook Files", The Wall Street Journal, October 1, 2021, https://www.wsj.com/articles/the-facebook-files-11631713039

29 ” A Meta engineer saw his own child face harassment on Instagram. Now, he’s testifying before Congress”, Barbara Ortutay, AP News, November 7, 2023, https://apnews.com/article/social-media-teens-meta-instagram-arturo-bejar-5f7fb7d55fb9f0da12cf3a57837fa0c5

30 ” 7 Ways Meta is Harming Kids: Findings from the Company’s Internal Studies”, Center for Countering Digital Hate, February 2024, https://counterhate.com/blog/7-ways-meta-is-harming-kids-findings-from-the-metas-internal-research/

31 ” His Job Was to Make Instagram Safe for Teens. His 14-Year-Old Showed Him What the App Was Really Like.”, Jeff Horwitz, The Wall Street Journal, November 2, 2023, https://www.wsj.com/tech/instagram-facebook-teens-harassment-safety-5d991be1

32 ” Community Standards Enforcement Report, Second Quarter 2021”, Guy Rosen, Meta, August 2021, https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021 /

33 “On Social Media, Transparency Reporting is Anything but Transparent”. Tech Policy Press. February 15, 2024. https://www.techpolicy.press/on-social-media-transparency-reporting-is-anything-but-transparent/

34 ” Transparency Center: Prevalence”, Meta. Accessed June 2, 2024. https://transparency.meta.com/policies/improving/prevalence-metric/

35 Code of Practice on Disinformation: A comparative analysis of the prevalence and sources of disinformation across major social media platforms in Poland, Slovakia, and Spain. TrustLab. September 2023. https://cdn.prod.website-files.com/661eb9d45168207d75d001c7/66563cf7a1f0730b04efe32f_Code-of-Practice-on-Disinformation-September-22-2023.pdf

36 ” Meta to Replace Widely Used Data Tool—and Largely Cut Off Reporter Access”. The Wall Street Journal, March 14, 2024, https://www.wsj.com/tech/meta-to-replace-widely-used-data-tooland-largely-cut-off-reporter-access-43fc3f9d

37 ” Twitter just closed the book on academic research”. The Verge, May 31, 2023. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

38 Exclusive: Elon Musk’s X restructuring curtails disinformation research, spurs legal fears. Reuters. November 6, 2023. https://www.reuters.com/technology/elon-musks-x-restructuring-curtails-disinformation-research-spurs-legal-fears-2023-11-06/

39 “Facebook is obstructing our work on disinformation. Other researchers could be next”. The Guardian. August 14, 2021. https://www.theguardian.com/technology/2021/aug/14/facebook-research-disinformation-politics

40 “Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn’t as Transparent or Accessible”. Columbia Journalism Review. July 9, 2024. https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php

41 Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017).

42 ” Snapchat makes its content guidelines public”. Axios Media Trends, March 15, 2023, https://www.axios.com/2023/03/15/snapchat-content-guidelines-public

43 ” Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process”, Meta, April 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards/

44 ” Who Moderates the Social Media Giants? A Call to End Outsourcing”, Paul Barrett, New York University Stern Center for Business and Human Rights, June 2020, https://www.stern.nyu.edu/experience-stern/faculty-research/who-moderates-social-media-giants-call-end-outsourcing

45 ” The Silent Partner Cleaning Up Facebook for $500 Million a Year”, Adam Satariano and Mike Isaac, The New York Times, October 2021, https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

46 ”The People Behind Meta’s Review Teams”, Meta, January 2022, https://transparency.meta.com/enforcement/detecting-violations/people-behind-our-review-teams/

47 “Why the internet’s most important law exists and how people are still getting it wrong”. The Verge. June 21, 2019. https://www.theverge.com/2019/6/21/18700605/section-230-internet-law-twenty-six-words-that-created-the-internet-jeff-kosseff-interview

48 “Opinion: Amend Section 230 to increase social media’s liability for drug sales on their platforms”. The Washington Post. October 27, 2021. https://www.washingtonpost.com/opinions/2021/10/27/amend-section-230-increase-social-medias-liability-drug-sales-their-platforms/

49 “Inside the Taylor Swift deepfake scandal: ‘Its men telling a powerful woman to get back in her box’ “. The Guardian. January 31, 2024. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

50 ” Herrick v. Grindr: Why Section 230 of the Communications Decency Act Must be Fixed”, Carrie Goldberg, Lawfare, August 2019, https://www.lawfaremedia.org/article/herrick-v-grindr-why-section-230-communications-decency-act-must-be-fixed

51 ” A Man Sent 1,000 Men Expecting Sex And Drugs To His Ex-Boyfriend Using Grindr, A Lawsuit Says”, Tyler Kingkade and Davey Alba, BuzzFeed News, January 2019, https://www.buzzfeednews.com/article/tylerkingkade/grindr-herrick-lawsuit-230-online-stalking

52 ” TikTok immune from lawsuit over girl’s death from ‘blackout challenge’ -judge”, Brendan Pierson, Reuters, October 2022, https://www.reuters.com/legal/tiktok-immune-lawsuit-over-girls-death-blackout-challenge-judge-2022-10-26/

53 ” ‘The tide has turned’: why parents are suing US social media firms after their children’s death”, Kari Paul, The Guardian, January 2024, https://www.theguardian.com/lifeandstyle/2024/jan/16/online-harms-social-media-lawsuits

54 ” Social media giants must face child safety lawsuits, judge rules”, Emma Roth, The Verge, November 2023, https://www.theverge.com/2023/11/14/23960956/meta-google-tiktok-snap-social-media-addiction-lawsuits

55 ”Reddit and YouTube must face a lawsuit over the radicalization of the Buffalo shooter”, Becky Sullivan, NPR, March 2024, https://www.npr.org/2024/03/19/1239478067/buffalo-shooting-reddit-youtube-lawsuit

56 ”Judge dismisses lawsuit claiming Amazon sold ‘suicide kits’ to teenagers”, Jonathan Stempel, Reuters, June 2023, https://www.reuters.com/legal/judge-dismisses-lawsuit-claiming-amazon-sold-suicide-kits-teenagers-2023-06-28/

57 ”Twitter dodges liability over tweeted child porn”, Eric Burkett, Courthouse News Service, May 2023, https://www.courthousenews.com/twitter-dodges-liability-over-tweeted-child-porn/

58 “Meta created a ‘Supreme Court’ for content. Then it threatened its funds”. The Washington Post. June 30, 2024. https://www.washingtonpost.com/technology/2024/06/30/meta-facebook-content-moderation-oversight-board/

59 ” Musk’s Twitter has dissolved its Trust and Safety Council”, The Associated Press, NPR, December 2022, https://www.npr.org/2022/12/12/1142399312/twitter-trust-and-safety-council-elon-musk

60 ”Twitter dissolves Trust and Safety Council”, Cat Zakrzewski, Joseph Menn, Naomi Nix, The Washington Post, December 2022, https://www.washingtonpost.com/technology/2022/12/12/musk-twitter-harass-yoel-roth/

61 ” What Is the Facebook Oversight Board?”, Cecilia Kang, The New York Times, May 2021, https://www.nytimes.com/2021/05/05/technology/What-Is-the-Facebook-Oversight-Board.html;
”Inside the Making of Facebook’s Supreme Court”, Kate Klonick, The New Yorker, February 2021, https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court

62 “Meta’s Oversight Board is too slow to matter”. The Verge. August 30, 2024. https://www.theverge.com/23852016/meta-facebook-oversight-board-too-slow-cambodia

63 ” Meta’s Oversight Board and the Need for a New Theory of Online Speech”, Paul Barrett, Lawfare, November 2023, https://www.lawfaremedia.org/article/meta-s-oversight-board-and-the-need-for-a-new-theory-of-online-speech

64 ” Meta rejects own board’s request to suspend account of Cambodian strongman”, Regine Cabato, The Washington Post, August 30, 2023, https://www.washingtonpost.com/world/2023/08/30/meta-cambodia-facebook-hun-sen/

65 ” Oversight Board recommendations”, Meta, May 2024, https://transparency.meta.com/oversight/oversight-board-recommendations/

66 ” The Meta Oversight Board and the Empty Promise of Legitimacy”, Evelyn Douek, Harvard Journal of Law & Technology, September 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4565180; ”Facebook’s oversight board blew up in its face”, Nicolás Rivero, Quartz, May 2021, https://qz.com/2005514/the-facebook-oversight-boards-trump-decision-blew-up-in-its-face

67 “The Internet Will Not Break: Denying Bad Actors § 230 Immunity”, Danielle Keats Citron, Benjamin Wittes, Fordham Law Review, https://ir.lawnet.fordham.edu/flr/vol86/iss2/3/

68 “Data Protection: Children’s Privacy”. Electronic Privacy Information Center. Accessed May 30, 2024.https://epic.org/issues/data-protection/childrens-privacy/

69 S.1418 - Children and Teens’ Online Privacy Protection Act. 118th Congress. Accessed May 30, 2024. https://www.congress.gov/bill/118th-congress/senate-bill/1418/text; S.1409 - Kids Online Safety Act. 118th Congress. Accessed May 30, 2024. https://www.congress.gov/bill/118th-congress/senate-bill/1409/text

70 US State Privacy Legislation Tracker 2024: Comprehensive Consumer Privacy Bills”. IAPP. Accessed May 29, 2024. https://iapp.org/media/pdf/resource_center/State_Comp_Privacy_Law_Chart.pdf; “Silicon Valley Battles States Over New Online Safety Laws for Children”. The New York Times. January 31,2024. https://www.nytimes.com/2024/01/31/technology/social-media-free-speech-netchoice.html

71 “Reps. Trahan, Schiff & Casten Introduce Digital Services Oversight and Safety Act”, Justin Hendrix, Tech Policy Press, February 23, 2022, https://www.techpolicy.press/reps-trahan-schiff-casten-introduce-digital-services-oversight-and-safety-act/
“Platform Accountability and Transparency Act Reintroduced in Senate”. Tech Policy Press, June 8, 2023, https://www.techpolicy.press/platform-accountability-and-transparency-act-reintroduced-in-senate/

72 ” What the FTC Does”, Federal Trade Commission. Accessed May 29, 2024. https://www.ftc.gov/news-events/media-resources/what-ftc-does

73 ” FTC’s $5 billion Facebook settlement: Record-breaking and history-making”, Lesley Fair, Federal Trade Commission Business Blog, July 2019, https://www.ftc.gov/business-guidance/blog/2019/07/ftcs-5-billion-facebook-settlement-record-breaking-and-history-making

74 “The Gaps Left Unfilled by the Senate Tech CEO Hearing on Child Safety”. Tech Policy Press. February 1, 2024. https://www.techpolicy.press/the-gaps-left-unfilled-by-the-senate-tech-ceo-hearing-on-child-safety/

75 ”Online Safety Act: Explainer”, Department for Science, Innovation & Technology, United Kingdom. Accessed 19 July 2024. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

76 "Online Safety Bill Volume 737: debated on Tuesday 12 September 2023". House of Commons. Accessed July 10, 2024. https://hansard.parliament.uk/Commons/2023-09-12/debates/81853BB7-375E-45C0-8C9D-4169AC36DD12/OnlineSafetyBill?highlight=%22safety%20by%20design%22#contribution-DFC3AD23-2095-4CC4-B8A6-467A80790B44

77 "Online Safety Bill: government amendments at Lords report stage", Department for Science, Innovation & Technology, United Kingdom. 30 June 2023. https://www.gov.uk/government/publications/online-safety-bill-government-amendments-at-lords-report-stage/online-safety-bill-government-amendments-at-lords-report-stage

78 Online Safety Act 2023, c. 110. Accessed 19 July 2024. Available at: https://www.legislation.gov.uk/ukpga/2023/50/pdfs/ukpga_20230050_en.pdf

79 Regulation (EU) 2022/2065 (Digital Services Act). Accessed 19 July 2024. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

80 Regulation (EU) 2022/2065 (Digital Services Act) Article 25. Accessed 19 July 2024. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065. "Providers of online platforms shall not design, organise or operate their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions"

81 See for example the most recent DSA transparency reports from Meta. Accessed 19 July 2024. Available at https://transparency.meta.com/reports/regulatory-transparency-reports/

82 The European Centre for Algorithmic Transparency (ECAT). 'About Page'. Accessed 17 July 2024.https://algorithmic-transparency.ec.europa.eu/about_en.

83 Regulation (EU) 2022/2065 (Digital Services Act) Article 40. Accessed 17 July 2024.Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

84 "Commission sends preliminary findings to X for breach of the Digital Services Act". European Commission. Accessed July 19, 2024. https://ec.europa.eu/commission/presscorner/detail/en/IP_24_376

85 "Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act". European Commission. Accessed July 19, 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373

86 "Commission opens proceedings against TikTok under the DSA regarding the launch of TikTok Lite in France and Spain, and communicates its intention to suspend the reward programme in the EU". European Commission. Accessed July 19, 2024. https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2227

87 "About us: Regulatory Schemes". eSafety Commissioner, Australian Government. Accessed July 2, 2024. https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes

88 "Responses to transparency notices". eSafety Commissioner, Australian Government. Accessed July 2, 2024. https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices

Building a Safe and Accountable Internet:
CCDH's Refreshed STAR Framework

Published September 2024