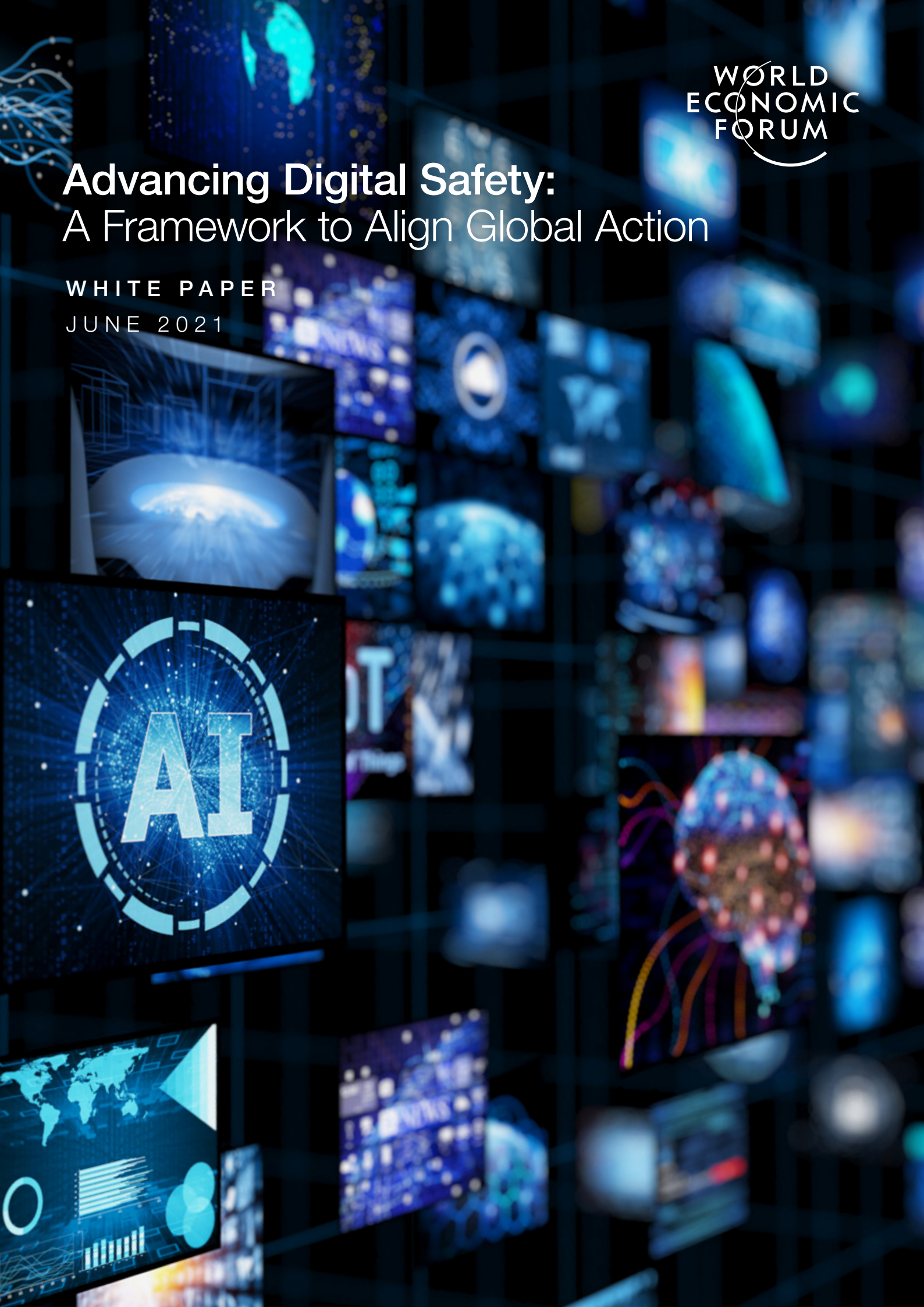
WORLD ECONOMIC FORUM

Advancing Digital Safety:
A Framework to Align Global Action

WHITE PAPER

JUNE 2021



Cover: GettyImages/Vertigo3d

Inside: Unsplash/nasa; Unsplash/KrzysztofKowalik; Gettyimages/DKosig; Gettyimages/Ridofranz; Gettyimages/gremlin; Unsplash/MikeKononov; Gettyimages/Eoneren

Contents

Executive summary

A user-centric framework establishing a safety baseline coupled with a regulatory framework to govern its enforcement can help mitigate exposure to harms online.

The past year and a half has put a spotlight on the role of digital platforms. The pandemic created immense challenges for countering misinformation on COVID-19 and vaccines. Social platforms set new precedents for curating content to promote scientific and authoritative sources, yet false information about the virus gained in speed and reach. In the United States, the 6 January Capitol insurrection necessitated a deeper look into the relationship between social platforms and extremist activity. During the pandemic, child sexual exploitation and abuse material (CSEAM) activity has increased according to INTERPOL[1] and Europol[2] reports.

The World Economic Forum has led the Advancing Global Digital Content Safety initiative since September 2019 to better understand the role of content practices, regulation and business dynamics in improving online safety from a user perspective. Key questions addressed include:

- How should the safety of digital platforms be assessed?
- What is the responsibility of the private and public sectors?
- What new regulatory schemes are needed? How should content liability laws be updated, if at all, to enhance safety?
- How can industry-wide progress be measured?

A majority (75%) of the experts surveyed agree or strongly agree that too much harmful content meets inadequate action on the part of platforms.

As many as 90% of those surveyed believe that content with a less clear definition of harm is tackled somewhat or highly ineffectively. Aligning on clear and consistent definitions of harmful content in these areas is a priority.

Through consultation with over 50 experts from academia, civil society, government and business, the Forum explored these questions through a mix of one-on-one interviews, workshops and a survey.

What became clear is that harm is a principle embedded in different national, regional and international legal frameworks and is a moral category that is context and culturally dependent.[3] While its definition is a matter of public concern, private industry action is needed to create and support safe online environments.

In industries such as transportation or energy, adherence to clear safety standards is required. Digital platforms that enable social communications should also have baseline safety thresholds, industry-wide standards and protocols, which do not currently exist.[4]

This White Paper distils a user-centric framework for advancing digital safety. Tensions between privacy, free expression, innovation, business incentives, private power and safety are all explored. Deficiencies are highlighted in thresholds for meaningful protection, auditable recommendation systems, complaint protocols and the use of personal details to minimize harm. A framework that answers the following questions is proposed as a path forward to crafting solutions that enhance user safety:

1. How much harm am I exposed to within this product?
2. Does this product have an undue influence over me or people I care for?
3. What avenues of remedy – public or private – are available if I am harmed?
4. Which details about me are being shared or exposed, and are they safe?

Industry standards that establish a safety baseline together with a regulatory framework to govern enforcement can help better protect users online. Collaboration across the public and private sectors must be urgently accelerated to counter false health narratives, deter coordinated acts of violence, and better protect children and adults online. It is vital that such collaboration be rooted in international human rights law, with a focus on protecting all rights for disadvantaged and marginalized communities. The UN Guiding Principles on Business and Human Rights provide a unifying framework on which to build.[5] The Forum's newly launched Global Coalition for Digital Safety will drive closer collaboration on solutions in this area.

1

The global challenge of digital safety

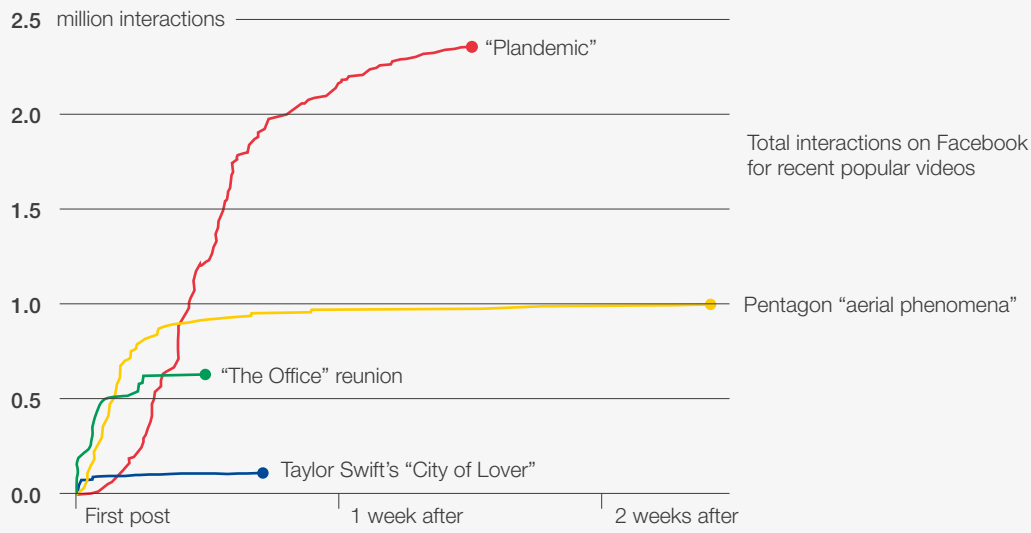
Health misinformation, violent extremism and terrorism, and child exploitation are areas requiring urgent attention.

1.1 Impacts of a global lockdown to online safety

While the challenges of online safety are not new, COVID-19 has brought several into focus. Specifically, health misinformation, violent extremism and terrorism, and child exploitation are areas in which content, conduct, contact and contracts have led to acts of harm or potential harm.[6]

Health misinformation. Addressing health misinformation has been a top priority for many public health officials, given the risk of illness or injury from inaccurate information and the documented speed at which false information spreads.[7] A recent Axios-Ipsos poll showing that the misinformed are less likely to get vaccinated highlights the urgency.[8] Many platforms have taken specific actions to combat false information on COVID-19, including disabling fake accounts and partnering with fact-checkers to label misinformation.[9] Platforms have also worked to remove problematic content. For example, YouTube has removed more than 800,000 videos containing coronavirus misinformation since February 2020, and specifically updated its COVID-19 policy in October 2020 to tackle vaccination misinformation.[10] YouTube also highlights that computers detect 94% of problematic videos before they are even viewed.[11]

Figure 1: Spread of the "Plandemic" movie online



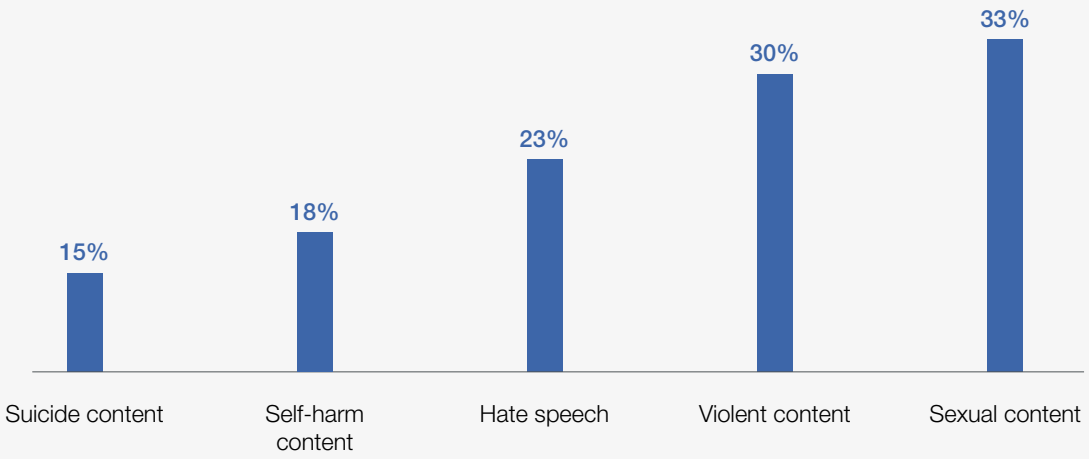
Source: "How the 'Plandemic' Movie and Its Falsehoods Spread Widely Online", *The New York Times*, 20 May 2020, https://www.nytimes.com/2020/05/20/technology/plandemic-movie-youtube-facebook-coronavirus.html (accessed 31 May 2021)

Depending on the platform, content with COVID-19 misinformation may be proportionately small: Facebook and Instagram removed over 1 million pieces of COVID-19 related misinformation[12] in the last quarter of 2020 considered to potentially cause "imminent harm", while an Oxford Internet Institute report found that less than 1% of all YouTube videos on coronavirus during its period of study contained misinformation. Yet those same YouTube videos were shared almost 20 million times, more than shares gained by the five largest English-language news sources. The spread of videos like "Plandemic" (Figure 1), a movie that promotes false information and conspiracy theories about COVID-19, showcases this disproportionate reach and speed.[13]

Violent extremism and terrorism. The storming of the US Capitol on 6 January and past attacks, including in Christchurch, have surfaced the mobilization and coordination of violent content, communities and actions on social platforms. Groups like QAnon and the Proud Boys gathered and organized not just on apps popular with alt-right groups, like Parler and Gab, but also on mainstream platforms.[14] While some social platforms took actions to remove efforts like the Stop the Steal movement on their technologies, for many, these actions seemed too little, too late. They also highlighted a potential gap in violent extremist content. For example, the Global Internet Forum to Counter Terrorism (GIFCT)'s recent call for taxonomic improvements suggests that current data collection and analysis practices may not have focused enough on domestic actors.[15] To others, platform responses were an example of overreach by private companies on matters of public concern; German Chancellor Angela Merkel openly criticized the banning of the former US President from Twitter.[16] With some US judges now banning Capitol riot suspects from the internet, questions of government overreach have also arisen.

Child exploitation. Children make up a significant portion of internet users – one in three is under 18 years old – equal to approximately 33% of children worldwide. As children engage in a wide array of online activities, they are likely to encounter certain risks (Figure 2).[17] One survey found that 70% of children have experienced at least one cyber threat.[18]

Figure 2: Percentage of children exposed to online risks



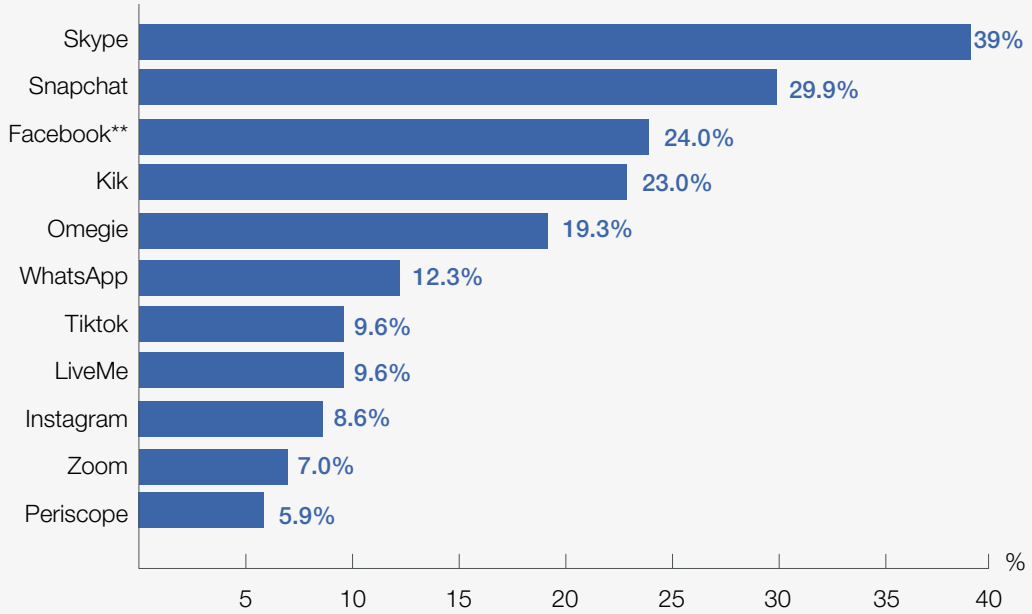
Note: Averages were calculated based on data from Albania, Bulgaria, Chile, Ghana, Italy, Philippines, South Africa and Uruguay.

Source: Based on Figure 26 in UNICEF, *Global Kids Online Comparative Report*, November 2019, p. 51

CSEAM consumption and distribution is growing despite international consensus that it is illegal. Channels, such as livestreaming, are being exploited to this end (Figure 3). More children are being "groomed"– that is, perpetrators use tactics and create conditions through the internet to sexually exploit children – resulting in an increase in overall self-generated CSEAM.[19] In the United States, complaints of child sexual abuse material and grooming have gone up by 75-200% depending on the state.[20]

Figure 3: Percentage of respondents who named specific apps/platforms* with live-streamed child sexual abuse

Most frequently mentioned apps, platforms and technologies that police officers reported seeing in their investigations of live-streamed child sexual abuse.



* Other apps mentioned several times were Facetime, Yahoo Messenger, Twitter, Youtube, Discord, Oovoo, Wickr, Stickam, Younow, Viber, Chatroulette, Chaturbate, ChatAvenue/KidsChat and Chatstep.

** (incl. Facebook messenger and Facebook Live)

The respondents answered an open question and could provide many answers.

Source: NetClean, *NetClean Report 2019: A report about child sexual abuse crime*, 2020, p. 32, https://www.netclean.com/wp-content/uploads/sites/2/2017/06/Netclean_report_2019_spread.pdf (accessed 31 May 2021)

Dissecting this increase is important: Facebook notes that more than 90% of the content reported to the National Center for Missing & Exploited Children (NCMEC) in October and November 2020 was the same or visually similar to previous reports. Copies of just six videos were responsible for over half of the child exploitative content reported during this period.[21]

A focus on prevention, and not just detection and reporting, is necessary to stop re-victimization.

While initiatives such as the WePROTECT Global Alliance and the Technology Coalition are working to address such issues, more must be done.

1.2 Balancing fundamental rights and addressing challenging trade-offs

Several goals must be balanced with online safety:

1. Privacy and safety

In many cases, providing users with greater privacy control enhances safety. For example, allowing users on social platforms to set their profile to private can protect them from unwanted access. In this way, privacy is a mechanism for safety. However, privacy can also complicate safety. Recent changes to the European Commission's e-privacy directive requiring stricter restrictions on the privacy of message data demonstrate an unintended consequence of stronger legislation. For example, when Facebook stopped scanning its messages in response to the new regulation, referrals for CSEAM coming from the European Union fell by 46% in the first three weeks of this change.[22] Others in the industry, including Google and Microsoft, interpret the law differently and continue to scan for CSEAM. Since this time, the EU has finalized new temporary legislation to detect the sexual exploitation of children online.[23]

Another technology that experts agree is vital to privacy is end-to-end encryption (E2EE). However, detecting illegal material by proactively scanning, monitoring and filtering user content currently cannot work with encryption. The NCMEC estimates that 70% of Facebook's CSEAM reports could be lost with the roll-out of E2EE.[24] Acknowledging this in public interviews, Facebook CEO Mark Zuckerberg voiced his commitment "to build the safety systems to do as well as we can within the framework of an encrypted system before we roll out end-to-end encryption".[25] A UNICEF report states the necessity of E2EE for privacy and security while noting its significant drawbacks in identifying, investigating and prosecuting CSEAM.[26] It notes that technical and legal solutions that consider the proportional impact on rights are needed for all users. The split response in the World Economic Forum's expert community survey regarding the impact of modifying encryption policies echoes the need for continued deliberation.[27]

2. Free expression and safety

Similar challenges exist when it comes to the freedom of expression and opinion.[28] Some assert that Facebook, Twitter and other companies go too far in their content removal practices. But in the United States, for example, these private companies are not obligated to protect First Amendment speech rights and can moderate certain categories of harmful but legal ("lawful but awful") content.[29] As private establishments, each platform can set its own terms and policies as long as it abides by the laws in countries of operation.

Whether in the private or public realm, many human rights experts point out that speech should not impede on the human rights of others.[30] Experts pointed out that in many situations, targeted harassment is designed to silence or cause victims to self-censor. Therefore, unabridged speech without regard to harm can actually suppress speech, particularly for vulnerable groups.

On platforms where both adults and children are allowed, this can be difficult. Contrast the obligations from the International Covenant on Civil and Political Rights (ICCPR) with the United Nations Convention on the Rights of the Child (UNCRC).[31] For adults, the right to full freedom of opinion and expression exists within the larger right of self-determination, to "freely determine their political status and freely pursue their economic, social and cultural development".[32] Conversely, children do not yet have the ability to determine themselves. Instead, they are "entitled to special care and assistance", and adults are charged with protecting a child's future right to self-determination.[33]

In the United States, many platforms do not permit users under 13 years of age to use their services in order to comply with the Children's Online Privacy Protection Act.[34] Yet despite platform efforts to comply, 82% of children aged between 8 and 12 have profiles on social media and messaging apps, according to research from CyberSafeKids.[35] Further, harms could also exist on services specifically targeting children as well as from passive consumption (without the need for an account). Platform decisions and policies related to content are even more crucial, given the intended and unintended exposure to children.[36]

3. Liability, innovation and safety

Over four-fifths (81%) of the experts surveyed believe that publishers/content creators should have primary liability; 37% say sites like Facebook and YouTube should have only secondary liability for content on their sites.

In the United States, the prospect of repeal of Section 230 of the Communications Decency Act and other efforts to increase platform liability have received much attention as a way of curbing digital harm.[37] Experts consulted had varied opinions as to the proposed reforms of Section 230. Susan Ness, Distinguished Fellow at the German Marshall Fund of the United States and former Commissioner

> **When an oppressed minority seeks equality and justice, and freedom from the harm and violence brought on by the systematically privileged speech of others, that's not censorship, that's accountability.**
>
> **Malkia Devich-Cyril**, Founder, MediaJustice, USA

of the Federal Communications Commission, cautions that some legislative proposals to repeal or revamp Section 230 could have unintended consequences. When threatened by a potential barrage of lawsuits, platforms may be incentivized to overly block problematic but legal content, thereby chilling speech, or to drop hosting user-generated content entirely, or to refrain from voluntarily moderating problematic content, thereby allowing harmful material to remain online. While major platforms are able to absorb litigation costs, the toll on smaller platforms may be too great. Instead of unleashing litigation to drive corporate behaviour indirectly, it may be more effective to focus directly on legislating a framework of transparency with robust oversight and accountability. Other experts consulted, however, differ on the prospect of reform; they stated that immunity under Section 230 should be reserved for internet companies that are truly passive carriers.[38] Moreover, some experts believe that proportionate, necessary and legitimate liability measures would not burden smaller players and could boost innovation if carried out appropriately.

Determinations are difficult given the lack of available information – over 82% of the expert respondents indicated that the transparency of industry content moderation and detection practices is poor or very poor. Platforms continue to evolve in these areas; Facebook, for example, points to its quarterly *Community Standards Enforcement Report* and commitment to undergoing an independent audit as a means of providing transparency.

4. Business incentives and safety

The creation and distribution of online content is a big business. Within this ecosystem, much attention has been paid to the potential relationship between advertising-funded platforms and the type of content that proliferates. Platforms highlight that it is in their business interest to keep users safe so they return and continue engaging with the product; they note this is core to company success, regardless of the business model at play. Nevertheless, when experts were asked what measures would improve digital content safety, changes to the business model was by far their top selection, at 80% of respondents. In line with this response are consumer attitudes around advertising: globally, 66% of consumers say they avoid online ads whenever they can.[39]

Despite Facebook's efforts to curtail hate speech, including a civil rights audit, some of its largest advertisers boycotted the platform last year. Such action can only provide some monetary incentive to do better, since the top 100 advertisers make up less than 20% of Facebook's ad revenue.[40] Small to medium-sized businesses in fact drive the majority of platform revenue and these businesses can depend upon the reach of platforms to a broad and massive user base to achieve their marketing goals.

Some consulted experts noted that a tension between business incentives and safety exists because there are not enough competitors in the market to make safety improvements a true priority. The discussion on competition appears at the end of the paper.

5. Private power and public responsibility

Who decides what is harmful and what action should be taken to address digital harms has become more consequential given the number of people affected. Though some call for a larger governmental role, others highlight the risk of governments abusing the expanded power. Legislation requiring companies to respond to content takedown requests adds complexity to the shared responsibility between the public and private sectors. Germany's Network Enforcement Act (NetzDG) obliges social networks with over 2 million users to remove "manifestly unlawful content" within 24 hours of being reported, and has sparked similar legislation in India, Kenya, Malaysia, the Philippines, Russia, Turkey and Venezuela.[41] Google has highlighted that determinations of content illegality are among the most difficult for YouTube reviewers to make.[42] When legislation demands quicker action by the private sector, potential issues of accuracy and overreach regarding speech rights need to be considered, even if speed may be beneficial given the (often) immediate impact of harmful content.[43] Regardless of whether future decisions related to harmful content are more in the hands of the public or private sector, the underlying concentration of power requires checks and balances to ensure consistency, accountability and transparency in upholding human rights.

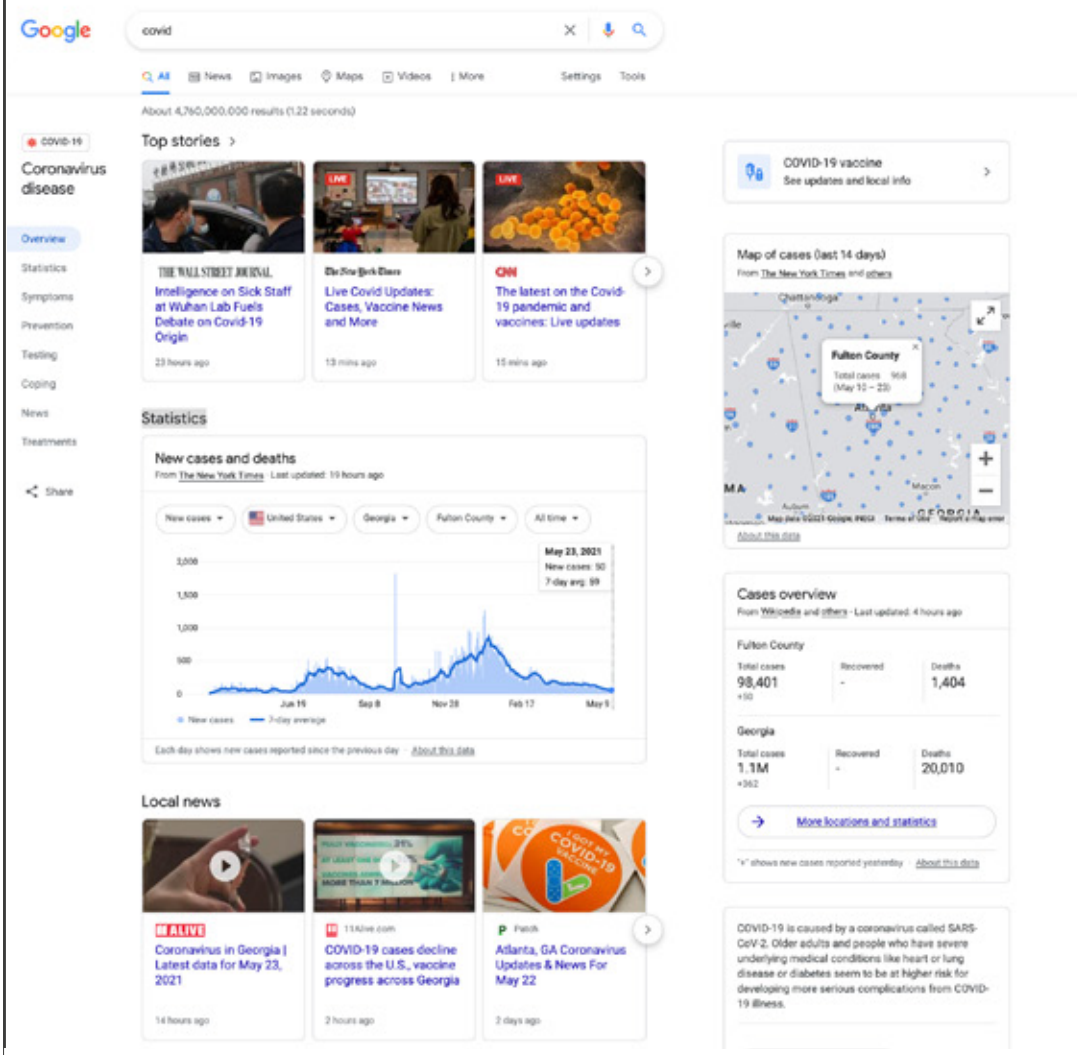
> **We don't have to start from scratch when making these decisions – we have a strong human rights framework and principles such as necessity, legality, proportionality to guide what action should be taken on harmful content. With the UN Guiding Principles on Business and Human Rights, we also have a framework for how responsibility should be shared between the public and private sectors.**
>
> **Lene Wendland**, Chief, Business and Human Rights, Office of the High Commissioner for Human Rights, Geneva

1.3 The complications of hybrid communication technology

Trade-offs immediately become more complicated with the "hybrid" nature of online communication, which crosses not only national borders but also governmental and industry regulation frameworks.[44]

Some online platform products hybridize more than others, and this mixing makes regulation difficult. Is Telegram, for example, a line of private communication – something akin to "common carrier" services such as phone calls and the post – or is it more like a town square? Technologically, it can be both. Should Facebook be regulated as a private community forum, a publishing platform (i.e. Facebook Pages), a broadcast service (i.e. Facebook Live) or an advertising service? Recent changes to Google Search also push the envelope: is it an index or directory, like the "Yellow Pages", or does it function as a news aggregator? The results of a recent Google search for "covid", for instance, produced top stories, local news and statistics with information from *The New York Times* (Figure 4). Each of these sociotechnical services within an online product offers different kinds of communication and thus different relationships with users. Safety requirements, therefore, also need to consider what the user seeks within these relationships and varied technologies.

Figure 4: Google search results for "covid", 24 May 2021



Source: Google Search, 24 May 2021

Before the internet, communications regulation followed the slower emergence of separate technologies within different countries. Locally determined rules exist for the postal service, broadcast radio or television, and legal advertising. Depending on the context, rules can be a blend of legislation and industry self-regulation. The film industry offers an example of the mix: Germany has a self-regulatory body for film ratings premised on a youth protection law, while Ireland has a statutory body connected to its Department of Justice, which examines and certifies all films distributed in the country. Government regulation can range from an advisory to a policing function depending on the country.[45]

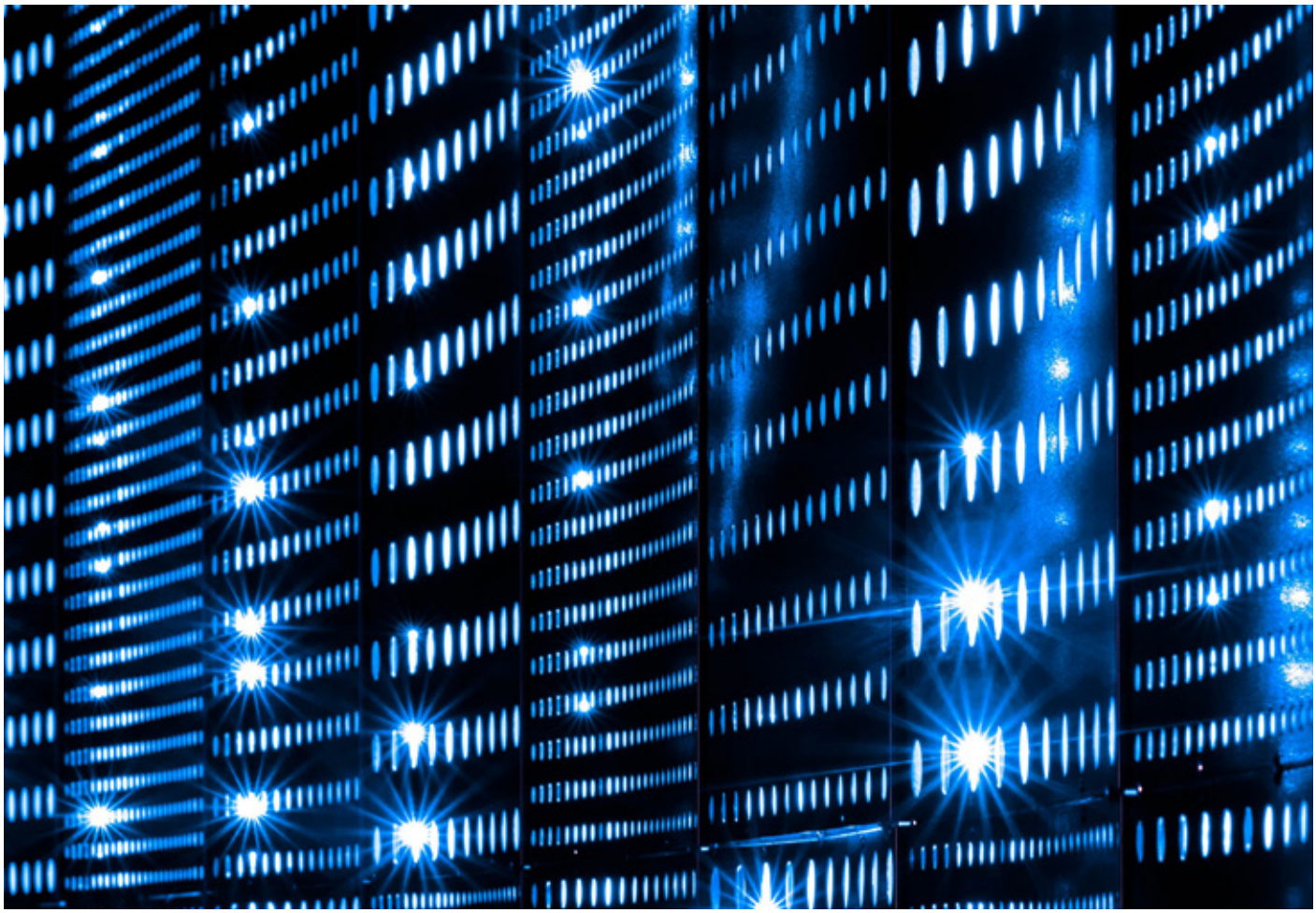
1.4 The difficulty of regulating

Regulating industry efforts to stem harmful content is not straightforward because of the difficulty in assigning responsibility and the potential unintended consequences of legal instruments.

One example is the illegal sharing of child sexual abuse photos. There is strong consensus that the perpetrators of these unlawful activities need to be held individually responsible, but pursuit is difficult. Many internet products do not require identity verification. Even when the identities of suspected criminals are discovered, the challenge of prosecution and extradition across regional and national borders can prove defeating.[46]

Given the difficulty in holding perpetrators accountable, one might attempt to hold hosting technologies better into account. However, this is challenging in terms of practicality and ultimate impact. Consider two acts – FOSTA (Fight Online Sex Trafficking Act) and SESTA (Stop Enabling Sex Traffickers Act) – that passed as exceptions to Section 230 in 2018. FOSTA-SESTA holds website publishers responsible if third parties are found to be "facilitating" prostitution on their platforms. Numerous websites took down ads and other parts of their sites, not because they were in violation of the law but because policing them was too burdensome, especially for the smaller platforms.[47] Had the bills been more focused on targeting websites known to facilitate sex trafficking, they may have been more successful in their ultimate quest.

> **We need globally aligned regulation that focuses not on the technology but on the systems and measures that demonstrate a safe platform. Companies must continue to invest in technology and partnerships to help combat harmful content online, and be held accountable to measures of prevalence, transparency and constant improvement.**
>
> **Simon Milner,** Vice-President, Public Policy, Asia-Pacific, Facebook, Singapore



2

The absence of a safety baseline enabling informed participation

Deficiencies in thresholds for meaningful protection, auditable recommendation systems, complaint protocols and the use of personal details are barriers to establishing a safety baseline.

2.1 The key challenge of safety

A key challenge before the world is: How can the risk of real-world harm stemming from online spaces be minimized for participants, and what actions are needed from both the public and private sectors to achieve this? Answering these questions requires an understanding of the following deficiencies in user safety:

2.2 Deficiencies in safety baselines

Many online platforms have increased their efforts to stem the tide of problematic content. Company transparency reports track some problem areas, and third-party efforts by Ranking Digital Rights, the Electronic Frontier Foundation and the Open Technology Institute, among others, continue to press for better results.[48] A new report launched as part of the Global Alliance for Responsible Media (GARM), a World Economic Forum flagship project, also offers a starting point for comparable safety metrics across platforms.[49]

Nevertheless, common understandings of what safety risks exist for participants are still not available.

Deficient thresholds for meaningful protection
When it comes to harmful content, there is currently no industry-wide accepted measure of user safety on digital platforms. Today, metrics reported on by platforms, which focus largely on the absolute number of pieces of content removed, do not provide an adequate measure of safety according to a user's experience; improvements in detection or the enforcement of action on harmful content according to platform policies, changes in these policies and categories of measure over time, and actual increases in the harmful content itself are not easily dissected. Even measures such as "prevalence", defined by one product as user views of harmful content as a proportion of all views, does not reflect the important nuance that certain groups – based on their gender, race, ethnicity and other factors – may be more exposed to harmful content.[50] Generally speaking, whether the large majority of content viewed on platforms is safe does not solve the problem for persons who are vulnerable.

The State of the World's Girls 2020 report by Plan International, which surveyed over 14,000 girls and young women in 31 countries, found that more than half of respondents had been harassed and abused online, and that one in four felt physically unsafe as a result.[51] UNICEF focus group research in East Asia highlighted that 40% of children had bad experiences online that they would not want to share or talk about to anyone.[52] GLAAD's Social Media Safety Index highlighted the issue for the LGBTQ community: 68% of lesbian, gay and bisexual (LGB) adults had encountered online hate and harassment.[53] Current platforms' efforts at assessment do not uncover such insights and could be complicated to measure, considering privacy and data security concerns.

Holistic measurement frameworks that go beyond the receiving end of content (e.g. consumption) to highlight the supply side of the information could help; metrics, such as the top 10,000 groups (based on members) by country or top 10,000 URLs shared with the number of impressions, could shed light on how, from where and by whom harmful content first originates.

Deficient standards for undue influence and use of personal information in recommender systems
The potential for the amplification of harmful content through recommendation engines (e.g. "suggested for you") is another area in which the standards for drawing the line between helpful, personalized suggestions and something akin to undue influence over users are deficient.[54] Part of this is due to a lack of understanding of the key inputs for these systems and any subsequent decisions about engagement optimization.[55] COVID-19 has highlighted several issues with, for example, Amazon removing certain products and directing customers to factual information about the disease. A recent audit of Amazon recommendation algorithms shows that 10.47% of search results related to vaccines promote "misinformative health products", which were also ranked higher than results for products that debunked these claims; clicks on a misinformative product tended to skew later results as well.[56] Overarchingly, it is unclear if and how problematic content and products are financially rewarded by current recommendation and advertising mechanisms, how this is linked to the use of personal information, and whether a conflict of interest exists regarding user safety.[57]

Deficient complaint protocols
Decisions regarding content removal, user suspension and other efforts at online remedy can be contentious. Depending on whom one asks, sometimes they may go too far or not far enough. Among many recent examples: YouTube's increased reliance on automatic removal during the lockdown caused too many videos to be removed, although it was an attempt to increase protection.[58] An adequate complaint response also includes the issue of timely and appropriate redress, a challenge given

the speed and volume at which content is created and distributed. Consider the response time of the Facebook Oversight Board regarding the platform's indefinite suspension of President Trump.[59] After direction from the Board in early May "to determine and justify a proportionate response" by early November 2021,[60] Facebook commuted the indefinite term to two years on 4 June.

Given the complexity of the case, this decision process may well be very efficient. However, there is no baseline on what is adequate within a company, let alone for an entity such as the Oversight Board. In other industries, customers of quality products are typically able to speak to a live company representative for further redress. When complaints are made internally to a platform, what constitutes a sufficient remedy process, and how accessible is it?

> **I think these companies need to increasingly shift the oversight over these kinds of decisions to external third parties that have the public interest in mind and command public trust.**
>
> **Dipayan Ghosh**, Co-Director, Digital Platforms & Democracy Project, Shorenstein Center on Media, Politics and Public Policy, Harvard University, USA

2.3 A user-centric framework for safety

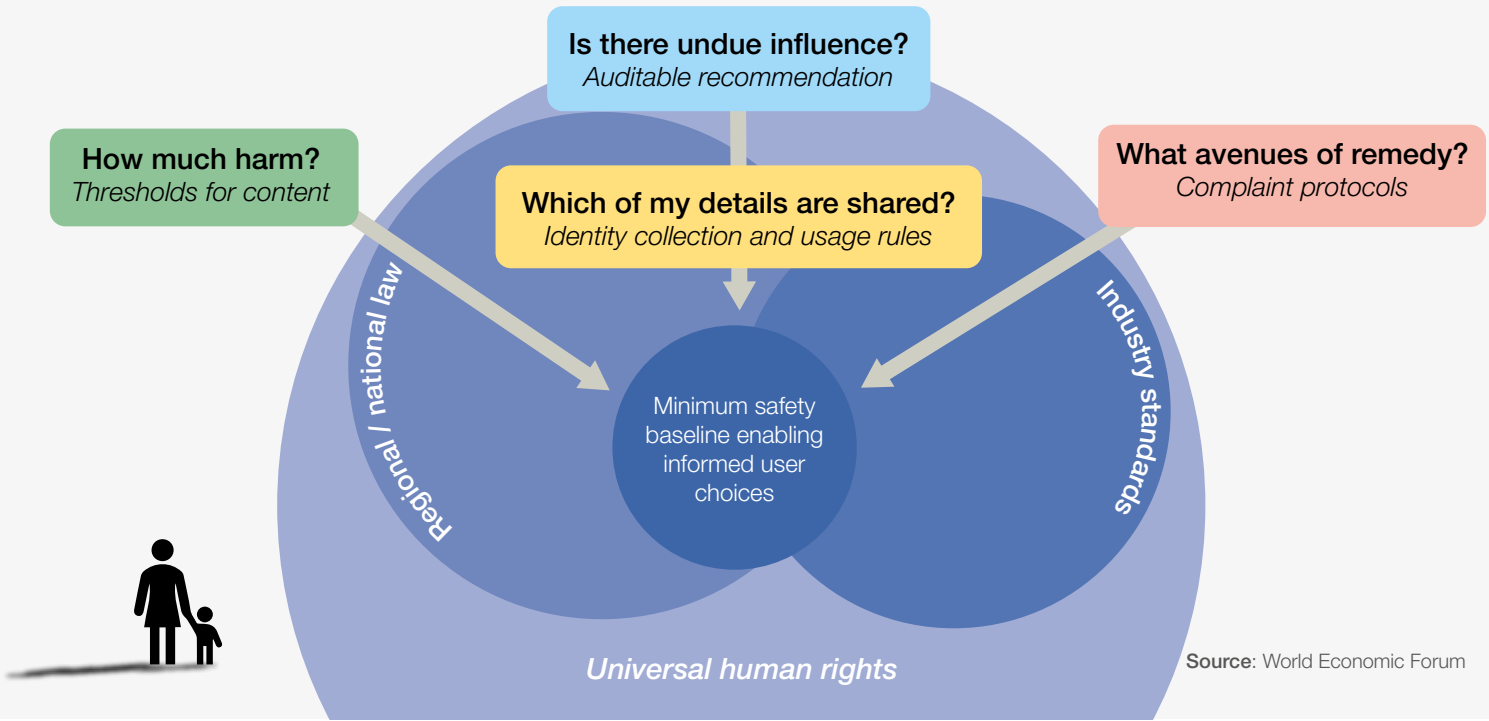
These deficiencies clarify that the challenge goes well beyond definitions of content and the limits of allowable speech. Making meaningful determinations about safe participation for ourselves and others by answering the following questions using the newly developed user-centric framework for safety (Figure 5) is the first step:

TABLE 1

Area	Safety information	Safety tension
Harm thresholds	How much harm am I exposed to within this product?	What is the line between free expression and speech that harms others?
Auditable recommendation systems	Does this product have an undue influence over me or people I care for?	What is the line between undue influence and tailoring to personal preference?
Complaint protocols	What avenues of remedy – public or private – are available if I am harmed?	What is the line between private versus public remediation?
Use of personal details	Which details about me are being shared or exposed, and are they safe?	What is the line between too little and too much personally identifiable information used by online platforms?

Answering these questions will require legislators, companies and participants to address key tensions with regard to people's rights and choices.

FIGURE 5 **A user-centric framework for safety**



Source: World Economic Forum

3

The need for public-private cooperation

Industry standards for a user-centric safety baseline that is enforced through regulation are needed moving forward.

3.1 Developing industry standards

Because safety is a determination negotiated in public understandings, it cannot be solved by any one company. Yet the development of safety baselines will need the participation of private corporations, since they know how harm unfolds in technological contexts and how to operationalize solutions.

Legal regulation around standards has been a necessary part of establishing trust in industrial systems, making them good both for the public and the market. Appropriate baselines require the development of shared professional and technical industry standards, different from individually defined approaches or commitments.[61] The Safety by Design principles developed by Australia's eSafety Commissioner in 2018, through multistakeholder collaborations, go some way to framing safety standards.[62]

> **You can't buy a car without working airbags or seatbelts – these design features are required and guided by international standards. The expectation of user safety should be as much a priority for technology companies as it is for the food, toy and automotive industries. Safety by Design should be seen as a business imperative and embedded into the values and operations of all digital platforms or services.**
>
> **Julie Inman Grant,** eSafety Commissioner, Australia

3.2 An ethical and fiduciary responsibility

In addition to external media regulations, ethical professional practices have been part of communications industries in ways central to today's concerns. Journalism ethics, for example, demand that practitioners not harm others through their storytelling, including careful reporting of suicide. Ethics in advertising acknowledges the need to be truthful in addition to respecting the legal boundaries established around public safety and child protection.[63] Future developments from organizations like the Trust & Safety Professional Association and the Digital Trust & Safety Partnership could be valuable to advance industry best practices.

The capacity for minimal safety through ethical standards often complements a legal approach to responsibility. Examples can be found in many professions, including the health (doctors), legal (lawyers) and financial sectors (accountants).

Fiduciary responsibility goes beyond voluntary standards to require various legal duties to beneficiaries, who include not only potential stockholders (as with owners of a hospital) but also the individuals receiving services (patient-as-client). With fiduciary responsibility, professional practitioners have a legal and moral responsibility to fulfil: *a duty of care*, which requires reporting to authorities when illegal harms are witnessed or suspected; *a duty of confidentiality*, which requires the protection of certain client data; and a *duty of loyalty*, which avoids conflicts of interest especially when guidance is provided.[64] Recently, the United Kingdom has focused on the duty of care in its consultations on the upcoming Online Safety Bill and the proposed work plan for the UK Digital Regulation Cooperation Forum.[65]

Differences in platform moderation approaches (i.e. centralized approaches taken by Facebook and YouTube compared to the community-based approach by companies such as Reddit, Wikipedia, Nextdoor or Discord) also need to factor into how this framework is applied appropriately. Further work is needed before practical application, considering the complexity of digital environments.

Sample questions

Threshold definitions that express the duty of care and confidentiality might include: How much is too much unsafe content within a product for the presence of children?

Auditable recommendation systems also pave the way towards fulfilling the duties of loyalty and confidentiality: Could companies' algorithmic system recommendations count as a kind of "expertise" bound by the duties of professional loyalty?

Adequate processes of complaint, to address issues of care and loyalty, might ask: How should platforms form appropriate Service Level Agreements for complaints resolution?

4 An agenda for action

In light of the need for more deliberate global collaboration, the World Economic Forum has launched a coalition for public-private cooperation to share best practices on improving digital content safety. Focus areas will include the balance of privacy, competition and safety in new regulation, cross-jurisdictional content cooperation, and global alignment on definitions of harm. Further suggestions for deliberation appear in the Appendix. Experts have suggested the following in the immediate term:

Stakeholder	Action area
Employers	Device protection and detection: Given 1 in 500 work computers are used to view CSEAM, employers might secure corporate mobile and laptop devices by utilizing leading tools that detect and restrict access to illegal content.[66] Employee training: Because some harmful content online may disproportionately target minority communities,[67] employers can ensure that acceptable conduct online through company platforms (e.g. Chatter, LinkedIn) is included in any existing diversity/inclusion training.
Digital platforms	Measuring safety (user experience): It is necessary to begin complementing current measures with metrics that focus on the user experience of safety and work to improve this over time; more informed choices by users and advertisers can be enabled across platforms. Cross-platform collaboration: A growing tactic used by bad actors is coordinated, cross-platform "volumetric attacks" to harass an individual or a group of individuals across multiple platforms. Online service providers should establish joint privacy-preserving mechanisms to counter this type of online harassment through better information sharing and policies.[68] Safety as a leadership priority: A Chief Risk Officer/Chief Safety Officer responsible for user safety on the platform should be designated to work with C-suite officers in all decision-making.[69] A more proactive approach to safety (e.g. eSafety's Safety by Design industry assessment tools, threat modelling, quality assurance processes) should be included. Content moderation workforce: Compensation, work environments, communication with content policy teams and support structures for moderators are currently unconducive to advancing digital safety; globally aligned content moderator workforce standards are necessary moving forward.[70] In particular, addressing issues highlighted by current workforce contracting models and the potential inadequacy of resources committed to content moderation require further action.[71] Peer community: Joining existing initiatives such as GARM is recommended, as is aligning with the structure in its aggregated measurement report.
Advertisers	Expenditure: Brand safety guidelines need updating to support safe user experiences, and spend must be allocated across platforms accordingly. Peer community: Joining existing initiatives such as GARM is recommended. Ad placement: More nuanced approaches to keyword usage and ad placement can be considered to fund positive content.[72]
Regulators	Monitoring and response: Outcomes of policy changes should be measured and adjusted accordingly (e.g. temporary derogation from certain provisions of Europe's ePrivacy directive to address child safety[73]). Online safety bodies: Consideration should be given to forming an office or body specific to online safety, such as Australia's Office of the eSafety Commissioner. Given the size, growth and importance of digital platforms to the lives of citizens, countries could consider establishing an independent statutory authority to help safeguard citizens by: – Preventing harm online through proactive awareness, research and education of citizens; – Protecting users by enabling them to report harmful content when there is inadequate redress through company channels; – Proactively driving systemic change by encouraging the upfront company adoption of safety principles in the design and product development process. Ex ante regulation: Given difficulty in reversing online harm once inflicted, focus is needed on ex ante regulation (e.g. EU) while remaining mindful of rights such as freedom of expression.[74] Peer models: Global cooperative models/frameworks that share best practices among countries (e.g. Agile Nations) should be given consideration.

5

Business incentives and market competition

Going above and beyond minimum safety baselines in digital products and services requires a deeper look into business dynamics.

With a baseline for user safety established, the business incentive to go above and beyond minimum thresholds becomes stronger. For example, a comparable safety score or rating (e.g. A++, 5-star system) across platforms that takes into account factors such as adherence to Safety by Design principles and enforcement effectiveness could drive further improvements. A safety score could focus more on quantifying positive language and interactions, thereby creating a differentiated value proposition for advertisers and consumers.

> **Safety should be viewed as a revenue-generating investment rather than a cost. Incentivizing competition based on safety can continue to raise the bar of what good looks like for safe online communities.**
>
> **Chris Priebe**, Founder and Executive Chairman, Two Hat Security, Canada

Another market-focused avenue for change is fostering an innovative and competitive media ecosystem. One in five girls, according to Plan International, has left or significantly reduced her use of a social media platform after being harassed, meaning that the majority still engage.[75] Currently, it is unclear if users on large social platforms who experience harm stay on these sites due to lack of alternatives or for other reasons, such as the network benefits of friends and family within the same product. Stronger competition can offer consumers and advertisers a wider array of choices for how they spend their time and money. Approaches to increasing competition might differ across markets: while the United States has focused on anti-trust actions, the United Kingdom has highlighted more specific interventions around data sharing and interoperability.[76] With more options for consumers and advertisers, safety could become a competitive differentiator; safety might fuel innovation in business models and the creation and deployment of new safety technology that reduces risks of harm.[77]

> **Without competition, dominant firms don't have to invent or innovate because their customers have nowhere else to go. The lack of competitive alternatives has allowed these companies to claim that this business model is the inevitable monetization choice.**
>
> **Rohit Chopra**, Commissioner, Federal Trade Commission, USA

6 Conclusion

Governments, industry, academia and civil society can drive collective action through the newly launched Global Coalition for Digital Safety.

How much harm am I exposed to? Is there undue influence? What remedy is available? Which personal data is used and how? These are questions that all users should be in a position to answer for themselves. The application of clearer safety thresholds, complaint protocols, auditable systems and identity guidelines can help both governments and industry – especially in technologies that transcend country boundaries – to consider their obligations to their citizens and users, respectively. The World Economic Forum invites governments, industry, academia and civil society to drive collective action through the newly launched Global Coalition for Digital Safety.

Appendix: Coalition considerations

Practical considerations for duties and responsibilities on social platforms

Internal

Company structure and operating model

- How is safety embedded in the company's business structure (e.g. is an executive on the leadership team)?
- Is there a dedicated safety team? Does the safety team work collaboratively and authoritatively with other teams to provide recommendations on risks and appropriate guard rails for new features?
- Is safety, like security, an integrated element of the product design and roll-out process?
- Regarding corporate decisions, what role does a safety or integrity team's view have on commercial decisions? How is safety considered in product/service testing?

Policies and practices

- Do clear, transparent and comprehensive policies on harmful content with accompanying rationales exist?
- Are moderation processes and technologies for detection, enforcement and complaints held to certain Service Level Agreements (SLAs) and standards for effectiveness?
- Are sufficient resources dedicated to content moderation?
- Are simple mechanisms for reporting content violations in place to inform users on both sides of the violation of what action was taken, if any, and why?
- Are users able to contact a live company support representative to resolve any escalated complaints?
- Are users able to "rate" their satisfaction with the resolution?

External

Mechanisms for accountability

- Are there independent third-party audits of content moderation practices/decisions, recommender systems, complaint protocols and use of personal data to monitor accuracy and effectiveness?
- Are independent third parties (e.g. external fact-checking organizations) used to support content decisions?
- Does external expert consultation or input shape the company structure according to human rights principles?
- Are mechanisms in place to support secure, privacy-maintaining data access for vetted academic or independent research?

Consistency with industry-wide standards

- Are clear measures of safety (content) available for users and advertisers, according to industry defined thresholds?
- Are SLAs in place to hold companies to appropriate detection, enforcement and resolution outcomes and time frames with users and advertisers?

Considerations for implementing standards

Regarding industry-wide technical or operational standards for safety, grey areas of harms definition illustrate how discussions might be advanced. Thoughtfully integrated media literacy approaches, according to a user-centric framework, are an example for Global Coalition action.[78]

1. *Harm thresholds.* Safety thresholds may depend on clear technical definitions. Yet cases of grooming are complicated for both humans and machines to assess since they consist of a set of textual and/or video exchanges over time.

   Defining the line between expression/opinion and illegal communications might include:

   - Clearly defined harm/potential harm in systems through shared taxonomies
   - Published thresholds for product-wide issues in children compared to adults
   - Related media literacy materials, including uncertainty or risk.

   Establishing a meaningful threshold may benefit from a conceptual or legal approach. Examples include:

   - Nuisance (e.g. noise, pollution): should an acceptable level of public-health-related misinformation be established?
   - Containment or contamination: Is an understanding of illegal material being used (e.g. CSEAM) that is like containment (e.g. nuclear regulation around radiation)? Or is it more like water and public utilities: in the case of incitement to violence, can randomized spot checks work?[79]

   Thresholds also rely upon data collection against standard definitions or taxonomies (e.g. NCMEC, GIFCT), with up-to-date classifications or categories.

2. *Auditable recommendation systems.* Consideration should be given to whether false advertising or incorrect information about health science has led to the purchase of products. Strong standards for accountable algorithms do not yet exist though principles are emerging.[80] Fields such as election observation and verification may offer auditing models to follow.

   Defining the line between manipulative influence and personal preference may be aided by:

   - Design documents and possible code review, available to accredited auditors
   - Ad libraries available for review, with metrics related to possible market reach
   - Quality assurance processes to test output, especially when content reaches a certain pace/reach
   - The auditing of system outputs by independent third parties
   - Related media literacy materials for children and adults.

3. *Complaint protocols.* When possible legal expressions of violence against elected leadership or candidates for office occur (e.g. “kill your senators”, “public execution” of specific individuals), what elements of a protocol can best negotiate the tension between private and public resolutions?[81]

   Defining the line between private and public remediation may be benefited through:

   - Required elements of internal complaint and appeals processes according to the platform governance structure (whether corporate or community defined)
   - Efficiency metrics, including the average speed of response and appeal
   - Reporting of complaint resolution and appeal processes (Santa Clara Principles)
   - Related media literacy materials for children and adults.

4. *Use of personal details*. Regulation related to the handling of private data exists: the European Union’s General Data Protection Regulation (GDPR) required many companies to change. Standards related to privacy, such as ISO 12859 and 30137, demonstrate how companies have been defining and implementing industry practices.

   Increased avenues for participants to know which personal details are shared in online services would be helpful, answering:

   - What is acceptable and necessary regarding personal information during product/service sign-up?
   - How is the anonymization/de-identification process handled, if relevant?
   - How is personally identifiable information handled with regard to the civil liberties of privacy and due process?
   - What protocols of sharing information securely exist across internal and external products (i.e. interoperability rules)?
   - Can enhanced media literacy materials help children and adults make more informed choices?

Contributors

World Economic Forum

Farah Lalani
Project Lead, Advancing Digital Content Safety, World Economic Forum

Cathy Li
Head of Media, Entertainment and Sport Industries, World Economic Forum

Lead author

Connie Moon Sehat
Director, News Quality Initiative, Hacks/Hackers, USA; Senior Project Fellow, World Economic Forum

The full list of contributors is available here: link

The World Economic Forum thanks Craig Newmark Philanthropies for its support, and David Bray, Inaugural Director, Atlantic Council GeoTech Center, USA for his collaboration on workshops under this initiative.

Endnotes

1. INTERPOL, *Threats and Trends: Child sexual exploitation and abuse – COVID-19 impact*, September 2020.
2. Europol, *Catching the virus: Cybercrime, disinformation and the COVID-19 pandemic*, 3 April 2020.
3. See Feinberg, Joel, *The Moral Limits of the Criminal Law, Volume 1: Harm to Others*, Oxford University Press, 1987.
4. Transatlantic Working Group, "Freedom and Accountability: A Transatlantic Framework for Moderating Speech Online", Annenberg Public Policy Center, June 2020; Halgand-Mishra, Delphine, et al., "Working Group on Infodemics Policy Framework", Forum on Information & Democracy, November 2020.
5. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework", United Nations, 2011.
6. Misinformation is "false or misleading information, spread unintentionally, that tends to confuse, influence, harm, mobilize, or demobilize an audience". See Spies, Samuel, "Defining 'Disinformation'", Version 1.1 updated with key points, MediaWell, Social Science Research Council, 29 April 2020; Livingstone, Sonia, and Mariya Stoilova, "Content, contact, conduct and contract – updating the 4Cs of online risk", Children Online: Research and Evidence (CO:RE), updated 8 March 2021.
7. Vosoughi, Soroush, Deb Roy and Sinan Aral, "The spread of true and false news online", *Science*, vol. 359, no. 6380, 9 March 2018, pp. 1146-1151; Ball, Philip, and Amy Maxmen, "The epic battle against coronavirus misinformation and conspiracy theories", *Nature*, vol. 581, 27 May 2020, pp. 371-374.
8. Nather, David, "Axios-Ipsos poll: The misinformed are less likely to get vaccinated", Axios, 1 April 2021.
9. Bras, Andrea, "How Platforms Curate and Elevate Reliable Information During Emergencies", NewsQ, 14 August 2020; Bursztyn, Leonardo, et al., "Misinformation During a Pandemic", Working Paper, Becker Friedman Institute for Economics, 1 September 2020.
10. BBC News, "YouTube deletes 30,000 vaccine misinfo videos", 12 March 2021.
11. Wakabayashi, Daisuke, "YouTube discloses the percentage of views going to videos that break its rules", *The New York Times*, 6 April 2021.
12. Wong, Queenie, "Facebook removed more than 1 million posts for COVID-19 misinformation", CNET, 11 February 2021.
13. United Nations, Department of Global Communications, "UN tackles 'infodemic' of misinformation and cybercrime in COVID-19 crisis", 31 March 2020; Wong, "Facebook removed more than 1 million posts for COVID-19 misinformation", op. cit.; Knuutila, Aleksi, et al., "COVID-Related Misinformation on YouTube", Oxford Internet Institute, COMPROP Data Memo 2020.6, 21 September 2020; Li, Heidi Oi-Yee, et al., "YouTube as a source of information on COVID-19: A pandemic of misinformation?", *BMJ Global Health*, vol. 5, 1 May 2020.
14. Wamsley, Laurel, "On Far-Right Websites, Plans to Storm Capitol Were Made in Plain Sight", NPR, 7 January 2021.
15. Global Internet Forum to Counter Terrorism (GIFCT), "GIFCT Launches Multi-Stakeholder Effort to Develop an Expanded Taxonomy Framework for the Hash-Sharing Database", Letter from the Executive Director, 24 February 2021; Lapowsky, Issie, "Tech spent years fighting foreign terrorists. Then came the Capitol riot.", Protocol, 8 March 2021.
16. Fillion, Stéphanie, "Angela Merkel Criticizes Trump's Twitter Eviction", *Forbes*, 11 January 2021.
17. Stoilova, Mariya, et al., "Investigating Risks and Opportunities for Children in a Digital World", UNICEF, Innocenti Discussion Paper 2020-03, February 2021.
18. Boston Consulting Group, "BCG Children Protection Cyberspace Survey", 2021 (unpublished).
19. Of recent and particular concern is a report by INTERPOL that found increases in online activity relating to CSEAM since COVID-19 started. See INTERPOL, "INTERPOL report highlights impact of COVID-19 on child sexual abuse", 7 September 2020; Kristof, Nicholas, "Opinion: The Children of Pornhub", *The New York Times*, 4 December 2020.
20. NetClean, *NetClean Report 2020: COVID-19 Impact*, 2021.
21. Davis, Antigone, "Preventing Child Exploitation on Our Apps", Facebook, 23 February 2021.
22. Hern, Alex, "Facebook under pressure to resume scanning messages for child abuse in EU", *The Guardian*, 20 January 2021.
23. European Parliament, "Provisional agreement on temporary rules to detect and remove online child abuse", Press Release, 30 April 2021.
24. National Society for the Prevention of Cruelty to Children (NSPCC), et al., Letter to Mark Zuckerberg "RE: Facebook's proposals to extend end-to-end encrypt messaging services", 6 February 2020.
25. Thompson, Nicholas, "Mark Zuckerberg on Facebook's Future and What Scares Him Most", *Wired*, 6 March 2019.

26. Kardefelt-Winther, Daniel, et al., “Encryption, Privacy and Children’s Right to Protection from Harm”, UNICEF, Innocenti Working Paper 2020-14, October 2020.
27. See Koomen, Maria, “The Encryption Debate in the European Union: 2021 Update”, Carnegie Endowment for International Peace, 31 March 2021, which highlights the main areas requiring additional deliberation.
28. Kaye, David, *Speech Police: The Global Struggle to Govern the Internet*, Columbia Global Reports, 2019.
29. Supreme Court of the United States, Manhattan Community Access Corp. et al. v. Halleck et al., No. 17-1702, 587 U.S. (2019).
30. For example, Nissenbaum, Helen, “Privacy as Contextual Integrity”, *Washington Law Review*, vol. 79, no. 1, 2004, pp. 119-157.
31. United Nations General Assembly, “International Covenant on Civil and Political Rights (ICCPR)”, 16 December 1966; United Nations General Assembly, “Convention on the Rights of the Child (UNCRC)”, 20 November 1989.
32. United Nations General Assembly, ICCPR.
33. United Nations General Assembly, UNCRC.
34. Jargon, Julie, “How 13 Became the Internet’s Age of Adulthood”, *The Wall Street Journal*, 18 June 2019.
35. Raidió Teilifís Éireann, RTÉ News, “Social media use among young children surges during lockdown”, updated 9 February 2021.
36. United Nations Office of the High Commissioner for Human Rights, “Committee on the Rights of the Child, General Comment on children’s rights in relation to the digital environment”, UNCRC General Comment No. 25, 2021.
37. Jeevanjee, Kiran, et al., “All the Ways Congress Wants to Change Section 230”, *Slate*, 23 March 2021.
38. Federal Trade Commission, “Tech Platforms, Content Creators, and Immunity”, Prepared Remarks of Federal Trade Commissioner Rohit Chopra, American Bar Association Annual Spring Meeting, 28 March 2019.
39. Ipsos & the Trust Project, “The Future of Trust in Media: Graphic Reference Report”, October 2020.
40. McCarthy, John, “Despite ad boycott, monopoly probe and a pandemic, Facebook revenue grows 11%”, The Drum, 31 July 2020.
41. Mchangama, Jacob, “Analysis: The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship”, Justitia, 5 November 2019.
42. See Google’s explanation in “Removals under the Network Enforcement Law”, Google Transparency Report, https://transparencyreport.google.com/netzdg/youtube?hl=en (accessed 1 June 2021).
43. See, for example, Carl Grossmann Verlag, “Liesching et al: Das NetzDG in der praktischen Anwendung” [in German], 26 February 2021.
44. McQuail, Denis, and Mark Deuze, *McQuail’s Media and Mass Communication Theory*, Sage Publishing, April 2020; Napoli, Philip, *Social Media and the Public Interest: Media Regulation in the Disinformation Age*, Columbia University Press, August 2019.
45. Burack, Cristina, “How Germany’s film age-rating system works”, DW, 7 December 2017; Irish Film Classification Office (IFCO), “What We Do”, Department of Justice, 2017; Ardana, Gilang, and Peter Sean Lie, “Guardians of the Screen”, American Chamber of Commerce in Indonesia, 6 March 2019.
46. Marks, Joseph, “The Cybersecurity 202: Democrats push a bill to combat child pornography without undermining encryption”, *The Washington Post*, 7 May 2020.
47. Romano, Aja, “A new law intended to curb sex trafficking threatens the future of the internet as we know it”, Vox, 2 July 2018.
48. Examples include the Electronic Frontier Foundation’s “Who Has Your Back?”, the Ranking Digital Rights Corporate Accountability Index, the Santa Clara Principles on Transparency and Accountability, and Open Technology Institute’s “Transparency Reporting Toolkit: Content Takedown Reporting”.
49. World Federation of Advertisers, *GARM Aggregated Measurement Report, Volume 1*, April 2021.
50. Australian Government, eSafety Commissioner, “Adults’ negative online experiences”, August 2020.
51. Goulds, Sharon, et al., *Free To Be Online? Girls’ and young women’s experiences of online harassment*, State of the World’s Girls, Plan International, 2020; World Wide Web Foundation, “The online crisis facing women and girls threatens global progress on gender equality”, Web Foundation, 12 March 2020.
52. UNICEF East Asia and the Pacific Regional Office and the Centre for Justice and Crime Prevention, *Our Lives Online: Use of social media by children and adolescents in East Asia*, UNICEF, 2020.
53. GLAAD, “Social Media Safety Index”, 2021, https://www.glaad.org/sites/default/files/images/2021-05/GLAAD%20SOCIAL%20MEDIA%20SAFETY%20INDEX_0.pdf (accessed 2 June 2021).
54. Undue influence may be difficult to demonstrate but is being considered by Martin Ebers, Lauren Scholz and Sarah Valentine. One example: Helberger, Natali, “Profiling and targeting consumers in the Internet of Things – A new challenge for consumer law”, in *Digital Revolution: Challenges for contract law in practice*, Nomos Verlagsgesellschaft, 2016, pp. 135-161.

55. See work by Sandvig, Christian, et al., "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms", Paper presented at the Data and Discrimination conference, Seattle, WA, 22 May 2014; Cobbe, Jennifer, and Jatinder Singh, "Regulating Recommending: Motivations, Considerations, and Principles", *European Journal of Law and Technology*, vol. 10, no. 3, 2019; Raji, Inioluwa Deborah, et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, Association for Computing Machinery, 2020, pp. 33-44.
56. Juneja, Prerna, and Tanushree Mitra, "Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation", Proceedings of the 2021 CHAI Conference on Human Factors in Computing Systems, Article 186, May 2021, pp. 1-27.
57. Australian Competition & Consumer Commission, *Digital advertising services inquiry: Interim report*, Commonwealth of Australia, December 2020.
58. Barker, Alex, and Hannah Murphy, "YouTube reverts to human moderators in fight against misinformation", *Financial Times*, 20 September 2020.
59. Patel, Nilay, "Facebook Oversight Board delays decision on Trump ban", The Verge, 16 April 2021.
60. Facebook Oversight Board, Case decision 2021-001-FB-FBR, 5 May 2021.
61. European Commission, Shaping Europe's digital future, "Code of Practice on Disinformation", 26 September 2018; World Wide Web Foundation, "Contract for the Web", November 2019, https://contractfortheweb.org (accessed 2 June 2021).
62. Australian Government, eSafety Commissioner, "Safety by Design", May 2019, https://www.esafety.gov.au/about-us/safety-by-design (accessed 2 April 2021).
63. Society of Professional Journalists, "SPJ Code of Ethics", 6 September 2014, https://www.spj.org/ethicscode.asp; American Marketing Association, "Codes of Conduct: AMA Statement of Ethics", https://www.ama.org/codes-of-conduct (both accessed 2 June 2021).
64. Possibilities and challenges are evolving in legal discussions. See Balkin, Jack, "Information Fiduciaries and the First Amendment", *UC Davis Law Review*, vol. 49, no. 4, April 2016; Khan, Lina, and David Pozen, "A Skeptical View of Information Fiduciaries", *Harvard Law Review*, vol. 133, no. 2, December 2019.
65. Government of the United Kingdom, UK Department for Digital, Culture, Media & Sport, "Online Harms White Paper: Full government response to the consultation", updated 15 December 2020; UK Competition & Markets Authority, "Digital Regulation Cooperation Forum: Plan of work for 2021 to 2022", Policy paper, 10 March 2021.
66. NetClean, *NetClean Report 2019: A report about child sexual abuse crime*, 2020.
67. United Nations Office of the High Commissioner for Human Rights, "Tsunami of hate and xenophobia targeting minorities must be tackled, says UN expert", United Nations, 15 March 2021.
68. One theoretical solution is Zero Knowledge Proof but implementation can be complicated.
69. This executive should collaborate closely with the Chief Data Officer/Data Protection Officer.
70. Roberts, Sarah T., *Behind the Screen: Content Moderation in the Shadows of Social Media*, Yale University Press, 2019.
71. Barrett, Paul M., *Who Moderates the Social Media Giants? A Call to End Outsourcing*, NYU Stern Center for Business and Human Rights, 8 June 2020.
72. GroupM, "Bluntly blocking COVID-19 keywords is not right: GroupM's Montgomery", Beet.TV, 30 March 2020.
73. Procedure EU 2020/0259(COD).
74. European Parliament, European Parliamentary Research Service, "Regulating digital gatekeepers: Background on the future digital markets act", Briefing, December 2020.
75. Goulds, et al., *Free To Be Online?* op cit.
76. UK Competition & Markets Authority, *Online platforms and digital advertising: Market study final report*, 1 July 2020.
77. York, Jillian, "Users, not tech executives, should decide what constitutes free speech online", *MIT Technology Review*, 9 January 2021.
78. Bustani, Camilla, "The Choice Challenge", International Institute of Communications, InterMEDIA, vol. 48, no. 3, October 2020; UNESCO, "UNESCO MIL Alliance", https://en.unesco.org/themes/media-and-information-literacy/gapmil (accessed 2 June 2021).
79. Froomkin, A. Michael, "Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements", *University of Illinois Law Review*, 1713, 2015; Ben-Shahar, Omri, "Data Pollution", *Journal of Legal Analysis*, vol.11, 2019, pp. 104-159; Rahman, K. Sabeel, "The New Utilities: Private Power, Social Infrastructure, and the Revival of the Public Utility Concept", *Cardozo Law Review*, vol. 39, 2018.
80. Diakopoulos, Nicholas, et al., "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms", https://www.fatml.org/resources/principles-for-accountable-algorithms (accessed 2 June 2021).
81. Hong, Nicole, "He Said to 'Kill Your Senators' in an Online Video. Now He's on Trial", *The New York Times*, 21 April 2021; Keller, Daphne, "Six Constitutional Hurdles for Platform Speech Regulation", Stanford Law School, Center for Internet and Society, 22 January 2021.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org