

ANNEX 1

SCALING TRUST ON THE WEB

CURRENT STATE OF TRUST AND SAFETY

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

The mission of the Digital Forensic Research Lab (DFRLab) is to identify, expose, and explain disinformation where and when it occurs using open-source research; to promote objective truth as a foundation of government for and by people; to protect democratic institutions and norms from those who would seek to undermine them in the digital engagement space; to create a new model of expertise adapted for impact and real-world results; and to forge digital resilience at a time when humans are more interconnected than at any point in history, by building the world's leading hub of digital forensic analysts tracking events in governance, technology, and security.

ISBN: 978-1-61977-279-3

This report is written and published in accordance with the Atlantic Council Policy on Intellectual Independence. The authors are solely responsible for its analysis and recommendations. The Atlantic Council and its donors do not determine, nor do they necessarily endorse or advocate for, any of this report's conclusions.

© **2023 The Atlantic Council of the United States.** All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Atlantic Council, except in the case of brief quotations in news articles, critical articles, or reviews.

Please direct inquiries to:

Atlantic Council
1030 15th Street, NW, 12th Floor
Washington, DC 20005

For more information, please visit www.AtlanticCouncil.org


June 2023



ANNEX 1**CURRENT STATE OF TRUST AND SAFETY****TABLE OF CONTENTS**

Introduction	2
The Evolution of T&S	3
Products and Platforms within the T&S Landscape	4
Key T&S Workstreams	6
Product Development	6
Policy Development and Enforcement	6
Tooling	7
Transparency and Accountability	7
Key Tradeoffs Within T&S	8
Protecting Rights vs. Mitigating Harm	8
Achieving Efficiency vs. Ensuring Accuracy	9
Ensuring Human Review vs. Depleting Human Resilience	10
Centralization vs. Decentralization	10
Growth vs. Safety	11
Short-Term Expenditure vs. Long-Term Value	11
Internal Process vs. External Expertise	12
Reactive Enforcement vs. Proactive Product Design	13
Looking Around: Key Sectors and Fields That Can Inform T&S Goals	13
Governance Models in Trust and Safety	15
Looking Ahead: Key T&S Challenges in Immersive Spaces	17
Content and Conduct Moderation	17
User Standards and Safety	18
Privacy	18
The Expansion of Apps and App Stores for XR	19
Equity and Access to XR Technology	19
Authorship and Acknowledgements	20

INTRODUCTION



Throughout its relatively short history, the practice of trust and safety (T&S) has undergone significant growth, yielding invaluable insights and the emergence of best practices. Lessons have been learned through both triumphs and challenges, leading to a deeper understanding of what actions should be taken and how they should be executed, and of frameworks that can help guide strategic thinking around these practices.

Robust collaboration and exchanges with peers in civil society and academia have played a pivotal role in shaping norms and standards, as exemplified by initiatives such as the [Santa Clara Principles](#), which have gone on to inform everything from regulatory strategies to the creation of entirely new civil society organizations.

Aspects of T&S that were once considered merely “nice to have” are now evolving into requisite, standard operating procedures throughout the technology industry. Regulatory pressure, established best practices, media attention, and compliance requirements with different parts of the technology stack have all contributed to establishing new prerequisites for how to operate technology platforms.

This applies to both large and small details—for example, appealing a content-moderation process was once a noteworthy service for users, offered by companies with additional resources or unique political will. Now, appeals have solidified their place within the Digital Services Act (DSA), signifying the evolution and increasing importance of robust T&S practices to doing business.

Despite the fact that T&S practices will play an instrumental role in shaping the landscape of technology in the twenty-first century, little is publicly documented about the field. Information about best practices and essential tooling remains trapped in silos, T&S teams inside companies are routinely embattled and under-resourced, and the many voluntary initiatives at the heart of T&S innovation are on [increasingly shaky ground](#). This is deeply troubling when the simultaneous emergence of T&S as a field is creating transformative new potential for collaboration, knowledge exchange, professionalization of T&S practices, and innovation across a range of stakeholder groups.

This annex seeks to give a light shape and context to the evolution of T&S and its workflows, the broad range of technologies that must incorporate T&S practices, the tradeoffs that practitioners and companies navigate

when considering T&S challenges, the diversity of governance models that have informed the development of the T&S field, and how existing T&S approaches may need to adapt in the race of immersive technologies.

THE EVOLUTION OF T&S

Rooted in the US technology sector, trust and safety emerged in the past fifteen years as a term to describe the teams and operations working to mitigate the harm (to users or others) arising from an online product or platform. This includes the use or misuse of the product, as well as negative interactions enabled, fostered, or intensified by the product's features that diminish trust among users of a product, or between users and the company offering the product.

Questions regarding trust, safety, and harm have existed since the earliest days of the internet. E-commerce, email, and online communities were beset almost immediately by fraud and spam; blogging and comment boxes immediately generated the need to counter the dissemination of child sexual abuse material (CSAM), hate speech, harassment, copyright infringement, and a wide range of other issues. As the internet shifted from individually produced content like websites and blogs to massive centralized platforms where millions—and then billions—of people could interact online, the scale of potential harms and negative social effects expanded dramatically and began to encompass understanding and responding not only to risks and harms facing individual users, but also to risks and harms occurring at societal levels.

No single definition of T&S holds across all audiences.¹ For some experts, T&S is “an umbrella term to describe the teams at internet companies and service providers that work to ensure users are protected from harmful and unwanted experiences.” For others, it is “the study of how people abuse the internet to cause real human harm, often using products the way they are designed to work.” For still others, it is the “the field and practices employed by digital services to manage content- and conduct- related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety. In practice, T&S work is typically composed of a variety of cross-disciplinary elements including defining policies, content moderation, rules enforcement and appeals, incident investigations, law enforcement responses, community management, and product support.”^t

In essence, T&S is an evolving term that gives shape to: a complex, dynamic array of policies, processes, tools, practices, and technologies that are deployed by individuals or teams (“practitioners”) inside of or working with tech companies, to keep the users of a particular online product or platform (or those affected by it) safe from harm, or at least reduce the likelihood, intensity, and frequency of harm.

Regardless of these variances, T&S remains a US tech industry-centric term that is only recently gaining greater traction as a field of study within academia as new initiatives strive to establish stronger academic underpinnings for the discipline. Policymakers and civil-society advocates, for example, use terms such as “platform accountability” or “platform governance” to frame concerns around the same harms that most companies would describe as falling within T&S. These include (but are certainly not limited to): hate speech, harassment, and defamation; misinformation and disinformation; child sexual abuse material and nonconsensual intimate imagery; terrorist or violent content; or trolling, brigading, and impersonation.

¹ This annex is meant to provide a broad framing of how T&S practices have been developed and operate within companies. Other sources cover this topic more thoroughly, and should be consulted for those seeking a deeper understanding of T&S. These include: *Introducing the T&S Curriculum*; *Digital Trust & Safety*, <https://alltechishuman.org/trust-and-safety-knowledge-hub>; and <https://datasociety.net/library/origins-of-trust-and-safety/>; *The End of the Golden Age of Tech Accountability - The Klonickles*; and <https://github.com/stanfordio/TeachingTrustSafety>.

² Please see DTSP's Glossary of Trust & Safety Terms for comprehensive definitions of common types of abuse addressed by T&S, common types of enforcement used in T&S, and common strategies for developing T&S solutions.

Today, T&S teams are grappling with some of the most consequential societal challenges around the world. It is increasingly clear that technology policy regulations being adopted or considered in multiple jurisdictions will only increase the need for qualified and resourced T&S teams and analytics, and that the promotion of healthy online communities will continue to prove a compelling value generator for companies with varied products, services, and business models. It is also increasingly seen as a “license to operate” function for companies engaged with user-generated content.

Finally, while T&S is now expanding globally as a field, it is important to note that the standards, practices, and technology that scaffold T&S were constructed overwhelmingly from US value sets. This US understanding of harms, risks, rights, and cultural norms has informed decades of quiet decision-making inside platforms with regard to non-US cultures and communities. Because its roots are so culturally specific to US and to corporate priorities, the emerging T&S field only represents one element of a much broader universe of actors and experts who also play a critical role in identifying and mitigating harm—including activists, researchers, academics, lawyers, and journalists.

As the T&S field develops into a range of different formal and informal structures, T&S practice opens up to a wider array of stakeholders.³ New channels for information exchange and learning exist in 2023 that can be game-changing for the dissemination of best practices and expertise both within the T&S practitioner community and between practitioners and a wider community of experts with aligned incentives in civil society, media, academia, and the public sector. Knowledge that was previously trapped within niche communities of practice inside large companies is finally seeing the light of day. This annex aims to illuminate a small part of that knowledge.

PRODUCTS AND PLATFORMS WITHIN THE T&S LANDSCAPE

Although it is common for discussions involving T&S to focus on social media platforms and content moderation, that tendency belies the wide range of products, platforms, conduct, and technologies covered by the emerging T&S field. It also underestimates the range and diversity of stakeholders who operate within the broader sector aimed at shaping T&S practices and principles. It is only in examining that larger ecosystem that systems-level challenges and opportunities begin to clarify, so there is tremendous value in narrating a range of the products and platforms that fall within the T&S orbit.⁴

Social media platforms create some of the most complex—or at least the most visible—T&S challenges, as their core aim is putting millions or billions of people in contact with each other. Part of this complexity arises from the platforms’ dependence on user-generated content to create revenue (through advertising, subscription, micro-payments, or some other monetization strategy). The largest platforms scale across many different vectors, simultaneously facilitating and fighting a variety of online harms across highly contextualized environments. Risks are particularly heightened when a platform begins to dominate a country’s media environment, telecommunications system, and business infrastructure.

Search engines facilitate and confront a complex range of harms, including: protecting users from malicious or fraudulent websites; combating the spread of misinformation and disinformation while balancing a range of contested facts; ensuring that search results do not include illegal or offensive content, while balancing rights to expression and to access information; supporting users who are under threat due to search results,

³ For a deeper analysis of how T&S is emerging as a field, see *Executive Report, Key Finding 1: The Emergence of a Trust and Safety Field Creates Important Opportunity*.

⁴ This analysis focuses on consumer-facing products; business-to-business products would also need to be considered in a more comprehensive systems-wide analysis.

and striving to ensure unbiased search results that do not discriminate based on factors such as race, ethnicity, gender, nationality, or religion.

Consumer-focused messaging applications such as Signal, Telegram, WhatsApp, Facebook Messenger, Google Chat, and others vary greatly across cultures and use cases, and continue to evolve as users' relationships with other online information-sharing spaces (like Facebook, Discord, Twitter, etc.) shift. Balancing privacy needs, encryption, and data-storage decisions with legal requirements to retain or disclose user data is an ongoing challenge within messaging. Just a sample of the issues that arise when chat apps are developed, or when they are embedded within other products, include: cyberbullying and harassment; scams and phishing attacks, the spread of spam, misinformation and disinformation, illegal content, CSAM, and violent or terrorist content; and protecting against the exploitation of children or the elderly. The central role that messaging plays in platform abuse adds an even greater level of complexity.

Streaming platforms can be broken into three primary groups: those like Netflix or Hulu that primarily provide access to official, licensed content; those such as Spotify or Apple Podcasts that allow for user-generated content but maintain licensed material; and purely user-generated streaming platforms like Twitch or YouTube. The T&S issues with the first group have focused primarily on advertising sensitivities (for example, whether to stream political ads) as well as rights to expression (such as Netflix's acquiescence to Saudi demands to block content). Regarding the second group, a longer list of risks exists based on the content being promoted by the service (which could include disinformation, incitement to violence, hate speech, etc.). The third group shares risks with social media companies and other user-generation-focused platforms, with the added and significant technical complexity inherent to moderating audio- and video-based content, particularly that which is livestreamed.

Gaming has long demonstrated key T&S concerns.⁵ Monitoring and policing problematic conversations between gamers can be challenging, especially because studies have shown that issues of sexism, racism, and extremist views are prevalent on gaming platforms. The large adolescent user base on many popular gaming platforms complicates these issues while also creating new policy needs, as do the complexities of moderating audio- and video-based content.

Dating apps must address harassment, hate speech, privacy risks, and geolocation risks, in addition to navigating the complexities of intimate image sharing, which may be consensual or nonconsensual—or illegal, in the case of minors or within local law. Given the additional sensitive nature of LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) issues, T&S teams at dating apps tend to focus on how platforms can create a safe space for all users.

Sharing-economy platforms such as Uber, Lyft, and AirBnB must confront a wide range of T&S issues, notably because of their role facilitating in-person interactions. Their teams focus on issues ranging from assaults, harassment, and hate speech to theft and fraud. They must also consider the physical safety of customers on both sides of the sharing arrangement, as well as the safety of individuals associated with the customer (for example, additional guests in a car or house) and damage to the physical components necessary for the sharing arrangement (e.g., a property or a car).

App stores—mainly Apple and Google—have come under increased scrutiny for which apps they allow to be on the platform and increased pressure to serve as front line defenders protecting users from malicious or unsafe applications. App stores play a monumental role in standardizing new T&S practices and influencing new norms. A range of increasingly popular app stores have emerged—including the Samsung Galaxy Store,

⁵ For a deeper analysis of the gaming industry, see [Annex 4: Deconstructing The Gaming Ecosystem](#).

Amazon Appstore, Steam, and Huawei AppGallery—while smart TVs, gaming platforms like Xbox and PlayStation, and streaming hardware also offer their own app stores.

Additional key products and platforms navigating T&S include [cloud-service providers](#), [content-delivery networks](#), [e-commerce platforms](#), advertising platforms (accompanying and monetizing some of the aforementioned [platforms and products](#)), “smart home” devices, wearables, transportation platforms, housing platforms, cryptocurrency, and video-conferencing/e-convening platforms.

KEY T&S WORKSTREAMS

Different teams “slice and dice” their T&S workstreams and core functions differently. The following is an illustrative tour of some approaches to T&S functions.

PRODUCT DEVELOPMENT

Similar to cybersecurity, it is a best practice to incorporate T&S objectives throughout the product-development process. This is necessary to ensure new products and features are designed with policies in mind, and to identify and address potential risks early in the product-development cycle.

- ▶ Designers play a crucial role in T&S efforts by seeking to create intuitive user interfaces that enable users to easily navigate safety features, such as managing privacy settings, blocking or muting other users, and reporting violative behavior.
- ▶ T&S personnel conduct risk assessments to explore ways in which a new product or feature could be subverted for fraudulent purposes or to harm users.
- ▶ Product managers and engineers translate T&S requirements into technical implementations, often applying safety-by-design principles as described above.

Integration may take the form of T&S teams working closely with product managers, engineers, designers, and other relevant teams. Additionally, some companies have dedicated T&S product managers, engineers, and other staff based in development teams.

POLICY DEVELOPMENT AND ENFORCEMENT

A foundational step for T&S teams is establishing and enforcing policies that communicate acceptable uses of a company’s products, as well as (where applicable) the types of content or behavior allowed on the platform. With each policy, companies typically develop a high-level version with which users and the public engage (some companies have referred to these as community standards or guidelines), and a more detailed internal version that companies utilize to enforce the policies. [The Trust and Safety Professional Association](#), a membership association for T&S professionals, [describes](#) several factors that influence how these policies are developed, including the mission and core audiences of a company, legal requirements, the opinions of consumers, and third-party or business partners such as advertisers.

Policy enforcement requires a substantial investment in product, policy, and operational support. Centralized strategies for detection and enforcement may take proactive and reactive measures, combining machine detection and user reporting with content review and investigations conducted by humans. Automated systems may play a key role in initial detection and enforcement, including proprietary artificial-intelligence (AI) models that aim to predict whether individual pieces of content are violative. Community-oriented models for detection and enforcement may offer users more features and functions for flagging low-quality content, or might give more authority to community moderators to adjudicate community-specific standards, even when those standards might be more stringent than the central platform’s. Companies with a high volume of users

often need to design complex systems and specific tooling to triage signals of potential policy violations (be it a user flag or an AI-generated alert) into a workflow for their T&S teams.

As companies have developed more holistic T&S regimes, enforcement mechanisms have advanced beyond binary “does it stay up, or get removed” decisions. Today, a piece of content that violates a policy could be removed, but it could also generate an array of other possible responses. For example, content could be allowed to stay on a platform with an added label or disclaimer suggesting that it might be misinformation, or content could be demoted, making it less discoverable to other users. How accurately and consistently such enforcement actions are taken remains at the core of discussions in T&S—certainly among observers who are critical of industry solutions to addressing harmful content. Faulty moderation of posts in non-English languages—particularly languages that are not dominant on the internet—and lack of agreement on edge cases (e.g., whether a post constitutes hate speech) are both commonly cited problems.

TOOLING

T&S requires a technical implementation layer that can become highly complex quite quickly, and is often built out over time with homegrown tooling suites and organizational structures as a company becomes aware of harms or risks. Effective T&S is as much a logistics challenge as a policy challenge—a matter of facilitating effective decision-making, undergirded by technology. T&S operations (which unite tooling and organizational workflows) can be thought of as an interactive looping through distinct goals.⁶ The central importance of tooling can be best illustrated by cases in which tooling is inadequate or absent. For example, companies’ internal systems are often not tailored for the needs of Global Majority users. Companies whose primary revenue-driving markets are English speaking and culturally Western have proven unlikely to invest in building high-quality classifiers for other markets and languages, even if their products have significant reach in those markets. The resulting poor T&S outcomes in these markets can often be attributed to a gap in appropriate tooling.

TRANSPARENCY AND ACCOUNTABILITY

The impact of content-moderation decisions on public discussion and life have made transparency reporting about the policy-development process, as well as the health of proactive and reactive moderation systems, an important pillar of T&S programs. Human-rights advocates and responsible business consultancies such as [Business for Social Responsibility](#) have advocated for more steps to publicly report on the development and efficacy of these systems. Transparency reports cover everything from enforcement around platform guidelines to government requests, and intellectual-property and human-rights impact assessments have [emerged](#) as another tool companies have used to address human-rights concerns related to content-moderation topics and product-development challenges.

Moves toward greater transparency have been welcomed by advocates, but it is important to note that [transparency-reporting](#) best practices and standards are still nascent, and involve complex tradeoffs regarding data storage, access, and retention. No shared understanding of transparency currently exists, nor does any clear basis for consistent comparisons company by company. Transparency reports can generate confusion or [facilitate obfuscation](#) in addition to increasing clarity. For example, if a given platform has publicly documented a high number of CSAM removals, does that mean the company is doing a superlative job of detecting and removing CSAM (i.e., good), or does it mean that the platform is awash in CSAM (i.e., bad)? Or does it mean that the company is more or less equal to other companies in terms of having, detecting, and removing the content, and simply has a better process for reporting its actions?

⁶ For a deeper analysis of this topic, see [Annex 2: Building Open Trust and Safety Tools](#).

Transparency should be seen as a key area for analysis and longitudinal study in the years to come, as more regulatory measures demand transparency reporting from companies in some fashion. In addition, clear and thoughtful transparency can be good for business by reducing accusations of bias, providing corrective guidance to users, and engendering trust. Notably, organizations like the [Digital Trust and Safety Partnership](#) are working to help define metrics and assessment tools that can be used across different companies to define whether a company's investments in T&S are adding value to the product, user community, and company, and where the company stands in relation to peer companies.

KEY TRADEOFFS WITHIN T&S

It is impossible to ensure that people can gather together in order to freely communicate, access information, engage with others, and build community, and also gather completely free from harm or risk. This is as true of offline spaces as it is online spaces, but the complexity and range of tradeoffs that must be navigated in online spaces are constantly evolving and consistently challenging. The following broad categories reflect some of the consistent tradeoffs that have helped to define core T&S work to date, and that will remain highly relevant as new technologies emerge.

PROTECTING RIGHTS VS. MITIGATING HARM

Balancing the protection of rights with the mitigation of harm has always been, and will remain, one of the biggest challenges impacting those charged with governing spaces where people come together online. Historically, rights to expression, association, access to information, and privacy have risen as the most critical rights that must be balanced—and, within the context of T&S, they are often balanced against user safety and also against brand safety. Further complicating this dynamic, “brand safety” may refer to reputational risk for advertisers generating revenue for a platform, or it may refer to the platform or hosting company's own reputational risk. Increasingly, companies are being called upon to balance broader and more diffuse societal forms of safety, managing harms that range from [widespread disinformation](#) to potential [addictions like gambling](#) that their platforms may propagate. No consensus exists on the clear basis or extent of any particular company's individual responsibility to consider user safety, let alone the safety of civic institutions or society itself, and that debate will help define the coming decade.

Where legal or normative consensus exists with regard to a particular type of harm, collaboration, technological innovation, and policymaking have arguably advanced more rapidly. For example, terrorism and child sexual exploitation and abuse both developed as early focus areas for T&S efforts—not only because the harm they cause is egregious, but because those behaviors were clearly criminal offline in most jurisdictions. This facilitated faster adoption of law-enforcement standards and governmental regulations at domestic and transnational levels; investments in staffing, product, and policy innovation within companies; and the creation of multistakeholder initiatives that could support knowledge sharing across a broad range of experts and stakeholders. These issues remain complex and contested spaces—and ones in which transparency is notably lacking in the quasi-government institutions partnering with tech companies—but the baseline normative agreements that were already in place allowed some coherence to develop more quickly than in areas such as hate speech or tech-enabled gender-based violence.

Indeed, one of the most dynamic and challenging issues that T&S addresses is the realm of “lawful but awful” content. This refers to content that is legal in a particular jurisdiction, but that is nevertheless considered unpleasant or harmful, especially when distributed at great scale, concentration, or velocity. What is “lawful” or “awful” varies between and within countries. What is legal in one state may be illegal in another (e.g., hate speech); what one community deplors may be uncontroversial elsewhere (e.g., blasphemy). And what may

be lawful for a person to express, may soon be unlawful for a company to recommend or amplify using an algorithm. Another critical current area of debate rests in balancing encryption and rights to privacy against legitimate law-enforcement and national security interests in monitoring and/or proving criminal activity.

One of the core ideas to prevent harm before technologies evolve is the concept of [safety by design](#), popularized by the Australian eSafety Commissioner [Julie Inman Grant](#). It focuses on embedding responsibility with the service provider around the content posted, accountability and transparency, and autonomy and empowerment of the user. The overall aim is to foster more positive, civil, and rewarding online experiences. Others have argued that centering users' human rights rather than their safety can achieve the same goals, but do so with greater respect for the range of tradeoffs required when balancing safety against other fundamental rights.

The extent of scholarship, research, effort, investment, and innovation that underlie the rights/harms tradeoff is too vast for any paper to attempt to summarize. Instead, it is critical to note that [any space](#) where humans gather online will require those governing—or seeking to govern—that space to balance safety against rights. As user numbers scale, so will the complexity of the tradeoffs that must be considered.⁷ This applies not only to governance policies, but also to how products and tools are developed and designed, and to how governing bodies—be they governments, multistakeholder forums, or companies—structure themselves.

ACHIEVING EFFICIENCY VS. ENSURING ACCURACY

Any online gathering space that allows user-generated content will inevitably need to balance the need to review content quickly with the need to review it accurately. Some form of automated content review (i.e., content moderation) will always be required, and some capacity to examine automated decisions for errors will also always be required. This is further complicated by the challenges of operationalizing certain policies at scale and the need for policies to adapt to newly emerging threats or environmental changes. No industry standard currently exists to guide companies in determining when and how to build in-house capacity for content review vs. when to outsource that capacity, and a vast gap exists between the capacity of the largest technology companies and that of almost all other companies. Indeed, smaller companies and start-ups may outsource the operational elements of content moderation, as well as aspects of their policy-development process.

Ensuring operational approaches—from workflow to tooling to company structures—consistently, accurately, and efficiently enforce policies is a substantial and continuously evolving effort that requires deep investment in technical and policy expertise and guidance. Even large companies quickly exceed their capacity to conduct human detection and review of contested or problematic content. To aid in detection and review, companies generally [invest](#) in some mix of building their own automated systems and artificial intelligence, and hiring external vendors that specialize in a particular type of analysis or review.

One key question will always be the accuracy of the automated system a company is using. Many companies utilize AI systems that are supervised models and require labeled data. It is difficult to develop robust models to identify and enforce against potentially violative content or behavior without a large dataset of previously identified violations. It is challenging to achieve meaningful enforcement in situations in which companies lack data, or sufficiently high-quality data, to be able to train their models. This helps to explain why the largest platforms are capable of building vast content-moderation systems, whereas smaller companies need to bring in external capacity.

⁷ For a deeper analysis of how greater interoperability between T&S and human rights could serve to strengthen both fields, see [Executive Report, Key Finding 4: Learning from Mature, Adjacent Fields Will Accelerate Progress](#).

Even the best algorithms, however adept they are at scanning massive amounts of flagged content (be it text, images, audio, or video), will miss nuance and still require humans to manually review content. Machines are trained to track predetermined pieces of language, data, or hashes, without context. Cultural and linguistic nuance remains a major challenge for even the biggest companies. Differential capacity to review content accurately across different languages or communities has been, and will remain, an evergreen problem that disproportionately exposes some communities to much greater risk than others. Humans who take on direct content review—in part to make automated decisions more accurate and more equitable—also take on proven, significant risk to their own physical and mental health. (See below for more on this topic.)

Technological innovations rapidly and consistently shift the accuracy and speed of automated approaches. The tooling and computing speeds necessary to support real-time audio- and video-based content moderation dramatically outpace the amount of audio- and video-based content users generate. Meanwhile, the introduction of generative AI capabilities to the general public will fundamentally shift standard practices for everything from creating content to reviewing it.⁸ What will not change is the fundamental challenge of balancing the need for accuracy with the need for efficiency.

ENSURING HUMAN REVIEW VS. DEPLETING HUMAN RESILIENCE

As noted above, human review of content is a widespread basis of T&S teams and practices, but that review comes at a cost. Indeed, T&S communities and organizations have grown rapidly since 2020, in part, because practitioners are seeking support from others who understand and sympathize with the challenges of their role. T&S practitioners—from frontline content moderators to individuals in public-facing leadership roles at large companies—take on T&S work at significant risk to their psychological, physical health, and, at times, physical safety. Working consistently at the heart of T&S dilemmas requires a level of resilience that most humans cannot sustain. This applies not only to T&S practitioners, but also to activists, researchers, and journalists, who often serve as first responders for their own constituencies. It is imperative that this truth be recognized, acknowledged, and addressed continuously as online spaces shift, evolve, and expand.⁹

CENTRALIZATION VS. DECENTRALIZATION

Centralized services have provided convenience and accelerated the maturity of the internet as we know it today. The increasing popularity of the fediverse raises new challenges, too.¹⁰ It is not yet clear how emerging regulations—such as, for example, the Digital Services Act (or traditional T&S knowhow)—will be applied to federated spaces. Standard T&S tooling relies heavily on centralized architectures, workflows, and data stores—none of which may exist in a federated space. Decentralized platforms operate without a central authority (or with central authorities having limited areas of scope), which means that individual administrators bear significant responsibility for creating or enforcing T&S policies or addressing harmful content. This opens up significant space for abuse or arbitrary decision-making in its own right. Decentralized platforms that allow pseudonyms or anonymous identities can incentivize expression by protecting rights to anonymity, while also making it harder to identify users who engage in harmful behavior. Absent a central authority, it can be difficult to enforce consequences for harmful behavior, such as banning a user, and it can also be

⁸ For more on how generative AI may be applied to content moderation, see [Annex 2: Building Open Trust and Safety Tools](#) and [Annex 6: Learning from Cybersecurity, Preparing for Generative AI](#)

⁹ For a deeper analysis of this topic, see [Executive Report, Key Finding 3: Protecting Healthy Online Spaces Requires Protecting the Individuals Who Defend Them.](#)

¹⁰ For a deeper analysis of this topic, see [Annex 5: Collective Security in a Federated World.](#)

difficult to challenge an unjust or ill-informed decision that punishes a user. This tension may prove to be increasingly important as immersive applications based more squarely in decentralized gaming architectures gain influence over gathering places online.

GROWTH VS. SAFETY

T&S professionals have pointed out the challenges companies face in balancing business goals, such as prioritizing growth (either user scale or revenue) with commitments to supporting healthy experiences on content platforms. T&S needs correlate closely with scale, but no bright line delineates where a particular element of growth (revenue, intentional expansion, adoption within new markets, etc.) should galvanize a proactive investment in new T&S policies, teams, services, or tooling in order to support user safety. Few resources currently exist to support companies or T&S teams in threat modeling particular scenarios and proactively mapping growth against safety (the equivalence of a heat map for societal unrest or division, for example). Without clear tooling in place, it may be unclear to those within a platform which languages are increasing in use or popularity on a platform, or which communities may be increasing their presence.

The absence of maturity models also continuously undermines T&S forecasting, investments, and prioritization. The T&S investments needed to mitigate a company's own reputational risk may not reflect the most endemic harms or risks on a platform, but rather one isolated incident of particular severity or one particularly controversial decision. A rare study of content-moderation costs for start-ups and mid-sized online service providers found that, for mid-sized companies, "cross-company collaborations following controversial or high-profile moderation decisions and could represent up to 10,000 work hours annually, the full cost of which [was] difficult to estimate given the varying salaries and opportunity costs implicated." This challenge continues today, even with new and well-resourced platforms.

SHORT-TERM EXPENDITURE VS. LONG-TERM VALUE

Making the business case for T&S is an ongoing and evolving conversation that should involve all company stakeholders.¹¹ When growth is the metric that a company's investors need to see before they will continue investing, the growth team—and the metrics it uses—will be center stage in orienting a company's efforts. This affects T&S work in several ways. First, the traditional imperatives of early and mid-stage tech companies—growth and revenue—drive the mission. T&S, both functionally and culturally, is often viewed as a side-show or speedbump to these key drivers or, worse, in fundamental tension with them. This dynamic often traces back to the early days of a company. A company is not lacking a T&S function, or even competent T&S work happening within the company. Rather, T&S concerns are ignored, discounted, or outweighed by the perceived needs of rapid growth and revenue, and categorized as something to prioritize later, once growth and monetization have been sorted out.

Even where company employees and executives value T&S, establishing constructive metrics remains a pernicious challenge. As with a municipality contending with public safety and well-being, T&S governance and enforcement are greater than the sum of their parts. Countless decisions on policy categories and individual cases have a collective effect on the health of an online space. There are some areas in which the overlap

¹¹ With a few noteworthy exceptions, the venture capital (VC) investors behind emerging technology either have not prioritized T&S issues or appear to be intentionally indifferent. In general, investors and executives have failed to connect durable value generation with investment in T&S practices. It is imperative to improve investors' understanding of the fundamental role T&S will play in generating value. Given the mad rush among VCs to fund AI-based products and companies, it will be critical for investors to understand where their AI investments would benefit from T&S teams or practices of their own, where AI-based approaches could actually further T&S, and what the limitations of AI are in a domain where human expertise and judgment have proven indispensable. For more on private investment as a market driver, see Executive Report, Key Finding 7: Systemic Harm is Driven by Market Failures that Must Be Addressed.

of user safety and traditional metrics are increasingly visible and aligned with robust T&S practices—such as user churn, where users who experience abuse or toxicity are more likely to stop using the product, sometimes very quickly. Similarly, a prevalence of scammy ads on a platform will decrease the clickthrough rate of all ads on the platform. However, many decisions are difficult to quantify.

If a company cannot measure T&S performance and impact, then incentives are difficult to align. At present, it is next to impossible for a chief operating officer (COO) or chief executive officer (CEO) to know if their T&S team is excelling or lagging. T&S is not amenable to conventional reporting metrics such as OKRs (objectives and key results), and requires a range of new metrics that capture the positive effects of T&S investments in a tangible way. Such metrics must tie into core product and engineering OKRs and metrics to ensure alignment across a company and, ideally, across tech sectors. If there are no solid metrics with which to measure safety, it's hard to make safety matter—hard to promote people based on effectively increasing safety, hard to orient teams around promoting safety, and hard to demonstrate to investors (when the company is privately held) or to investment analysts (if the company is publicly traded) that a product has achieved growth and revenue while also making meaningful advances in user safety.

The perception that T&S investments are a cost center, rather than a value generator, remains one of the greatest barriers blocking more widespread and consistent adoption of T&S practices and standards. This disconnect also fundamentally implicates how investors and boards consider T&S investments within broader parameters of due diligence and fiduciary duty. Mass layoffs in the T&S community in 2022 and 2023, as well as ongoing shifts in the structure and expertise companies are seeking as they take on heavier compliance responsibilities, have demonstrated how significantly externalities can impact T&S goals and strategies inside companies. Immense need exists to define stronger metrics and assessment tools that can be used across different companies to define whether a company's investments in T&S are adding value to the product, user community, and company, and where a company stands in relation to its peers. Some notable progress is being made in this regard. In addition, the emergence of new and widespread regulatory requirements will also fundamentally reshape how companies evaluate investments and forecast costs.

INTERNAL PROCESS VS. EXTERNAL EXPERTISE

Companies are not necessarily lacking outside stakeholders (some with useful subject-matter expertise, others mainly with political power) offering their opinions on what the company should do in any given situation, or as a matter of policy. Practical issues make it difficult to harness such subject-matter expertise when it is offered.

First, such subjects are almost never politically neutral. One safety issue for a given external stakeholder is often in tension with an issue close to the heart of another stakeholder—such as LGBTQ safety and rights on the one hand, and religious organizations protecting their adherents' right to free speech on the other, and the rights and safety of trans-exclusionary radical feminists on another. Each of these groups has different entities to advocate for its agency and to press companies to enforce terms of service “fairly”—often meaning in line with that group's worldview and values. These demands will sometimes be mutually exclusive.

Second, companies may rightfully be wary of too closely adopting the views or recommendations of any one organization at the risk of being seen as “rubber stamping” the values or preferences of any one particular outside organization, or of giving that organization (and, by extension, political partisans who support it) an inside track to having its moderation preferences implemented by the company.

Finally, standardized models for connecting external expertise to teams inside of companies— particularly T&S product and tooling teams—remain a significant and counterproductive gap within industry. This impacts expertise from civil society and academia.

- ▶ The onus continuously rests on civil society—which, as a field, comprises organizations that are generally smaller and less well resourced, and which navigate challenging operating environments—to adapt to the operational needs of well-funded, empowered corporations. Civil-society organizations lack insight into how the feedback they provide is used. Externally facing mechanisms focused on policy development or the reporting of “bad” content have been the most common mechanisms that companies have piloted, but they have not proven sustainable or effective, and can be perceived by civil society as token initiatives that pull precious time and focus, while offering limited impact in return.
- ▶ The state of practices, tools, systems, policies, and partnerships used in contemporary T&S practice is not captured in so-called transparency reporting mechanisms (reports, blog posts, etc.) by platforms, nor is it properly reflected in academic research. As a specific example, academics lack access to the same data sets and other information contained in companies, as well as the tooling that would allow them to analyze those data. Closing this gap is essential, as independent academic research helps accountability, innovation, and field-wide transparent dissemination of best practices.

REACTIVE ENFORCEMENT VS. PROACTIVE PRODUCT DESIGN

Many conventional and external understandings of T&S begin and end with enforcement—rules, policies, takedowns, timeouts, and account bans. For many years, T&S operations have revolved around enforcement, as well as intervention into the operation of the service. Teams of reviewers have relied on automation (sometimes extensively and other times more sparingly) to detect T&S violations and/or implement T&S decisions. But, an increasing part of T&S teams and their role within tech companies involves a more organic relationship with the product team—evaluating a potential or planned product or feature for the ways it is likely to be misused or abused, the types of harms that might be foreseen, and, in some cases, helping to figure out how to modify the product to mitigate those risks before it has shipped. This differs from traditional enforcement-centered work in a number of ways. It is proactive rather than reactive, and it is tied to the nature of the product itself rather than directed at intervening into human behavior by applying rules and policies. It leads to different staffing choices and focus areas for a T&S team—specifically, more people with experience in product, data science, and engineering. Increasingly, a modern T&S team is not just traffic cops, but seatbelt makers. These changes are still in flux and under way across the tech industry, but have deep implications for how product development is done, and the relationship among internal company stakeholders—product, engineering, user research, legal, T&S—collaborating on new product surface areas before they launch.

LOOKING AROUND: KEY SECTORS AND FIELDS THAT CAN INFORM T&S GOALS¹²

The technology sector has long suffered from the presumption that its problems are novel, and that relevant knowledge must then be developed *sui generis* in bespoke, tech-centric settings. Trust and safety arose through an attempt, in part, to address societal problems as they manifested in digital settings. The technology sector was late to recognize any larger responsibility to address those issues, which meant that other sectors have long been approaching similar questions from the other (non-technological) side of a problem.

¹² For a deeper analysis of this topic, see *Executive Report, Key Finding 2: Academia, Media, and Civil Society Bring Crucial Expertise to Building Better Online Spaces*.

T&S is just one subset of a much broader universe of actors from sectors such as academia, civil society, and media who have played critical roles in identifying and mitigating harm, even though they may not be seen (or see themselves) as operating within the T&S field.

The budding T&S academic initiatives described above (courses, journals, research conferences) are essential at a moment when the gap between practitioners and the academic community is large. More must be done to help ensure that practitioners are better informed by relevant academic research and, in turn, that academic research can be shaped by an accurate understanding of evolving practice. The current state of practices, tools, systems, policies, and partnerships used in contemporary T&S practice is not captured in so-called transparency-reporting mechanisms (reports, blog posts, etc.) by platforms, nor is it properly reflected in academic research. Closing this gap is essential, as independent academic research helps accountability, innovation, and field-wide transparent dissemination of best practices.

In addition to academia, civil-society organizations and independent researchers have always played critical roles in protecting the broader interests of T&S. Civil-society actors, especially in the Global Majority, have exposed the negative impacts of many platforms by identifying, naming, and analyzing harms or potential risks, including risks to human rights. Civil-society groups have also played a major role in analyzing the negative impacts of different revenue models and in bridging the gap between companies and high-risk or marginalized communities, especially through multistakeholder efforts.

Civil society also functions as a major lever for change. Groups have developed independent recommendations for the private sector, worked directly with individual platforms to provide counsel and expertise on complex questions involving their constituencies, and organized to shift political will at companies to respond to harms. The development of voluntary frameworks, such as the Santa Clara Principles and the Manila Principles, has helped drive forward debate and consensus around best practices and minimum acceptable standards for companies. Nongovernmental organizations (NGOs) have also fostered innovation by designing independent accountability frameworks and trackers, recommendations for product design, user interfaces, security features, reporting, and new features. Civil-society-driven work with marginalized communities has resulted in powerful new product offerings that have improved safely and driven growth.

However, standardized models for connecting external civil-society (and academic) expertise to teams inside of companies—particularly T&S product and tooling teams—remains a significant and counterproductive gap within industry. The onus continuously rests on civil society—which, as a field, comprises organizations that are generally smaller and less resourced, and which navigate challenging operating environments—to adapt to the operational needs of well-funded, empowered corporations. On top of this, civil-society organizations generally lack insight into how the feedback they provide is used. Externally facing mechanisms focused on policy development or the reporting of “bad” content have been the most common mechanisms that companies have piloted, but they have not proven sustainable or effective, and can be perceived by civil society as token initiatives that pull precious time and focus while offering limited impact in return.

Civil society can, and should, play an important role in proactive policy and system design. This would complement the capacities of professional T&S teams and deepen those teams’ understanding of issues like societal-level risks or specific bad actors. Civil society can also play a particularly important role in identifying how harms operate and evolve across platforms. This is an analysis that T&S teams inside companies often lack the access, resources, or permission to track themselves, but that is of critical importance to understanding and illuminating societal-level risks, as well as specific bad actors. Absent civil-society expertise, enormous gaps would open around the world in collective understanding of how harms propagate, and how products can be developed that protect fundamental rights and serve users’ needs.

Media have also played a key role in driving attention to T&S, notably in the areas of platform vulnerabilities. There are limitations and shortfalls within the current practice of technology journalism, though, as well as

threats to the future viability of independent media across the world. These include inattention to and ignorance of the issues among media professionals, a tech-industry backlash against investigative or critical reporting, downward pressures on journalism's business model globally and the subsequent hollowing out of newsrooms, and increasing political constraints on the free press across the world. Media coverage significantly shapes what the general public understands, whether or not that coverage is accurate or factual. Poorly reported or sensationalist stories exacerbate mistrust and rivalry between the tech industry and media. Additionally, the volume of poorly reported, technically inaccurate, or distorted coverage has real negative consequences for public understanding of technology, particularly when it comes to informing lawmakers and demand for regulation.

Significant value would be derived from improving relations between the sectors, including educating more journalists on relevant technical and policy issues, and engaging policy and product leaders within companies to better understand the role and value of the fourth estate. Increasing journalistic capacity to report on the impact of different platforms in marginalized communities, as well as across the Global Majority, is also key. Coverage of how platform decisions affect Global Majority countries is rarely at the front of the agenda, and the revelation of potential harms invariably comes after damage has been done.

GOVERNANCE MODELS IN TRUST AND SAFETY

“Who decides (and on what basis)?” is the existential question at the heart of T&S practice, as well as the industry in which it has developed. At the broadest level, no clear global law or normative framework applies to technology companies or (by extension) their T&S practice. The specifics of how internet companies are governed vary based on their size, business model, geographic location, and the prevailing legal and social contexts in which they operate. In addition to internal governance policies (as described above), companies may rely upon a wide range of additional governance models. The following offer an illustrative list of different approaches.

- 1 External engagement:** At the most micro level, companies may have internal-governance policies (as described above) that derive purely from a company's own values or priorities. Some companies, like Twitch, Meta, Spotify, or TikTok, may augment their governance structures or decision-making process by creating external (and generally non-binding) engagement mechanisms such as an advisory board or safety council. The Meta Oversight Board goes beyond that by operating as an independent entity, funded by Meta, with binding decision-making authority over isolated Meta cases. The board has hinted at grander aspirations.
- 2 Voluntary industry groups:** Other self-regulatory initiatives may go beyond a particular company to bring companies together in industry groups or associations that establish commitments to codes of conduct, principles, or principles. For example, the Oasis Consortium offers safety standards that companies can commit to uphold. The Digital Trust & Safety Partnership was founded by companies such as Discord, Google, LinkedIn, Meta, Microsoft, Patreon, Pinterest, Reddit, Shopify, Twitter, and Vimeo to share, develop, and promote industry best practices on issues related to trust and safety. Its Best Practices Framework aims to provide a uniform method to assess online content and conduct-related risks.
- 3 Voluntary multistakeholder initiatives:** Additional self-regulatory initiatives expand beyond industry to multistakeholder efforts that more closely resemble the multistakeholder models that have helped define internet governance. For example, the Global Network

Initiative (GNI) establishes voluntary principles to protect user rights to freedom of expression and privacy. Its corporate members commit to independent audits to ensure that they are in compliance with those principles, and work closely with academic, investor, and NGO constituencies inside GNI to develop principles, best practices, and knowledge exchange across the membership. The Global Internet Forum to Counter Terrorism (GIFCT) is a multistakeholder effort aimed at developing technological solutions against terrorism and violent extremism, conducting research, and sharing knowledge with smaller companies, as well as civil society and academia.

- 4 Government-created voluntary mechanisms:** Governments have also established voluntary mechanisms that can help move governance forward, even as they refrain from carrying the enforcement authority of regulations or legislation. The EU Code of Practice on Disinformation aims to motivate companies to collaborate on solutions to the problem of disinformation, and was strengthened by the European Commission in 2022. The European Commission has been clear that the Code of Practice, while voluntary, will become a central pillar of Digital Services Act compliance for platforms, significantly shifting the incentive structure in favor of this voluntary mechanism. The Christchurch Call: Home is a government-led, non-binding initiative to curb the spread of terrorist material online, launched by French President Emmanuel Macron and New Zealand Prime Minister Jacinda Ardern after the 2019 Christchurch terrorist attacks. The United Nations Guiding Principles on Business and Human Rights (UNGPR) increasingly inform assessments and voluntary principle mechanisms in the technology industry, and offer a set of guidelines for state actors and companies to prevent, address, and remedy human-rights abuses committed in business operations. However, these principles were not initially created to address human-rights risks in the online environment.
- 5 Civil-society-developed frameworks:** Companies may also sign onto principles that have been developed entirely outside industry. The Santa Clara Principles 2.0 emerged from a coalition of civil-society organizations and academics, as standards directed at state actors and internet platforms. The Manila Principles on Intermediary Liability were developed by several NGOs and digital-rights organizations, and serve as a roadmap for “internet intermediaries”—search engines, social networks, telecom companies, and internet services providers (ISPs). They offer a set of standards based on international human-rights instruments and other international legal frameworks to combat online censorship and other human-rights abuses.
- 6 Binding laws and regulations:** Finally, companies are subject to an increasing array of laws and regulations.¹³ Some, like the General Data Protection Regulation (GDPR), the Digital Millennium Copyright Act (DMCA), and Section 230 of the US Communications Decency Act (CDA), have long governed companies’ decisions regarding content and user data. The upcoming Digital Markets Act and Digital Services Act from the European Union are seen as once-in-a-generation laws that may fundamentally change how platforms operate and determine tradeoffs.

¹³ For a deeper analysis of the role of emerging regulation as a market driver, see *Executive Report, Key Finding 7: Systemic Harm Is Driven by Market Failures That Must Be Addressed*.

LOOKING AHEAD: KEY T&S CHALLENGES IN IMMERSIVE SPACES ¹⁴

All of the considerations raised in this annex will apply to emerging technologies including but not limited to the rapidly evolving world of extended-reality (XR) platforms, products, and tools. Many of the biggest issues in the XR ecosystem—content moderation, ads and monetization, user safety, privacy, sustainability, and access to technology—present similar manifestations of the challenges companies, regulators, and users have experienced in attempting to mitigate online expression and harm concerns on social media and internet platforms. Privacy and cybersecurity will be at play as well. For example, the volumes of data collected and traffic sent as part of gaming platforms are of interest to companies and governments—and, potentially, to criminal actors as well. XR environments may be centralized or decentralized, and the risks and opportunities present in those respective environments reflect those shared by non-XR spaces.¹⁵

One specific hallmark differentiating XR spaces from more traditional (or “flat”) spaces is XR’s focus on achieving fidelity, i.e., accurately reproducing or simulating the real-world environment, objects, or actions in order to make an XR experience look, feel, and sound as realistic as possible to a user. The neuroscience behind XR can lead to a blurring of what is or isn’t real; as a result, the consequences of harmful or inappropriate behavior may be more acute. Different levels of fidelity also impact the degree to which information about the user can be ascertained by their behavior within the ecosystem. In addition, the more that XR environments can create totally new scenarios and possibilities for users, the greater the possibility that new experiences in a virtual environment will create unforeseen harms.

Although this section is focused on charting harm and risk, it is critical to note that virtual-reality (VR) spaces can also create opportunities for unforeseen and uncharted benefits. For example, initial studies indicate that VR spaces can improve retention in educational programming or support individuals struggling with mental-health challenges, while innovative new VR-based initiatives are striving to increase awareness of human-rights violations or help prepare witnesses and victims to testify in international criminal tribunals.

CONTENT AND CONDUCT MODERATION

Many of the content-moderation issues discussed in the T&S space today (various incidents of bullying, harassment, hate speech, exposure to offensive/explicit/extremist content, dissemination of misinformation and disinformation) apply to XR as well. For example, groups and individuals have been found using games and game-related platforms to normalize extremist views, and survey-based research has demonstrated the continuing role that harassment plays in gaming environments online.

In addition to increasing the intensity of some harms, the richness and freedom afforded by higher-fidelity interactions and environments can also introduce new vectors for harm. Unlike in flat digital experiences, nonverbal cues such as facial expressions, eye contact, and body language are often made possible in VR. Compounding matters, such cues are often still difficult to interpret due to the current representational limits of technology and the absence of well-established social norms or codes of conduct, making it harder to accurately interpret the meaning behind someone’s words or actions. Current content-moderation norms and regulations (which are already complicated, fragmented, controversial, and quickly evolving) will have to be adapted to properly address the challenges presented in the XR ecosystem. In preparation for XR moderation, stakeholders will need to develop strategies for addressing familiar issues in new technological contexts.

¹⁴ For a deeper analysis of T&S considerations in XR, see [Annex 4: Deconstructing the Gaming Ecosystem](#).

¹⁵ For a deeper analysis of the specific challenges of responding to traditional T&S concerns in federated or decentralized spaces, see [Annex 5: Collective Security in a Federated World](#).

One major hurdle is the moderation of social VR and audio/chat functions. Similar to the challenges platforms with livestreaming face, moderating (whether manual or automated) this type of content is particularly difficult, and can be costly. Recently, moderation companies have invested in automated voice-chat moderation, while some are even exploring other forms of nonverbal and non-text-based moderation, though this remains particularly cost-ineffective. Of note, major gaming companies have announced recording voice chat for moderation purposes, and it is expected that more companies will follow suit soon. In a similar vein, when creating policies and terms of services to moderate users, companies will need to consider the unique ways in which users interact with technology that breaks the divide between virtual and physical worlds. This means adapting policy to focus on behavioral interactions in addition to speech-centric interaction, as well as developing tooling to support that shift.

As generative AI inevitably lowers the barrier to creating synthetic media, it is foreseeable that deepfakes and additional forms of audio- and video-based impersonation will increasingly enter XR spaces, creating new opportunities not only for harassment and disinformation, but also for financial fraud.

USER STANDARDS AND SAFETY

Widespread integration of XR will present new iterations of familiar challenges like harassment and problematic interactive media use. Though video-game and social media addiction have been more widely studied, other consumer-safety concerns have emerged in recent years, from eye strain to the psychological impacts of being physically or sexually assaulted in a virtual world. While the majority of VR headsets have traditionally been intended for those thirteen and older, early Food and Drug Administration (FDA)-approved VR treatments are aimed specifically at treating children with lazy eye. Specific risks to child safety will need to be considered and negotiated as adoption increases; indeed, Meta recently opened Horizons World to teen users in the United States and Canada, and placed specific limitations on their accounts. While child safety has historically been an easier issue on which to reach agreement (especially at the governmental level), different approaches are already being adopted to consider varying user experiences based on age. For example, some games contain design features intended to deceive or manipulate players (e.g., into playing longer, purchasing items), which might be considered harmful to vulnerable users (e.g., children).

Across all age groups, the adoption of XR technologies will force companies and stakeholders to explore and define consent, bystander notification, and user privacy (in a physical and virtual-bodily sense) as they pertain to immersive hardware. “Dark patterns”—where algorithms aggravate mental-health issues by proposing increasingly problematic or harmful content—also run the risk of being even more harmful in immersive environments. In addition, the normalization of chance-based monetization systems (sometimes called “gablification”) in games is raising important questions about T&S from both commercial exploitation and technologies specifically designed to foster compulsive behavior or even addiction among players.

PRIVACY

As with traditional social media, user privacy is critical. In the XR space, privacy is a combination of civil-liberties work, globally focused human-rights advocacy, gaming-related advocacy, and user-based online harms. The way privacy is conceptualized and ensured is different because of the increased interoperability inherent in the metaverse. Interoperability allows different virtual environments and platforms to communicate and interact with each other, but is also an increasing concern for the XR ecosystem. As XR hardware continues to evolve and standardize, user security and understanding of risks, opportunities, and assumptions of use will be important touchpoints for companies and regulators alike.

Increased computational capacity on devices also makes it more likely that individuals' phones will become their primary computers. This could mean that more data are being sent from phones. It could also mean that people will have a greater want for phone-based software applications than they had before. As companies and researchers experiment with using on-device computational capabilities, the evolution of privacy-preserving and machine-learning techniques, coupled with demands for more software services, will force policymakers to grapple with questions such as whether and how data can be protected; how much computation can realistically be used on mobile devices without rendering them ineffective or forcing users to ditch them for efficiency reasons; and if processing more user data on devices could risk companies waving away the risks of processing the data and generating insights from them. Companies may also use XR to capture more sensitive data on individuals, whether scans of a room from a VR headset or the sheer volume of privacy risks associated with eye-tracking technology and other forms of biometric data collection.

THE EXPANSION OF APPS AND APP STORES FOR XR

Augmented-reality (AR), VR, and mixed-reality (MR) app stores may increasingly play a role in this space as well. The Meta Oculus App Store, the SteamVR store, and other online marketplaces enable device users to install software on their headsets and interact, in different ways, with virtual worlds. Unlike in mobile app stores, which remain relatively concentrated in Apple and Google, AR/VR/MR app stores, at least for the time being, present consumers with more options—and developers have more places to create new software as well. In many ways, this reflects the merging of somewhat distinct, but deeply interconnected, connective industries with many of the AR/VR/MR platforms built upon long-standing gaming industry and players.

It is also worth noting that the Web3 ecosystem is already generating new business models such as decentralized marketplaces, where buyers and sellers can interact directly with each other without the need for intermediaries. This can lead to reduced transaction fees, increased competition, and greater transparency in the buying and selling process—but the same lack of intermediation may also raise new T&S challenges for effective monitoring and timely intervention, and exceed the capacity of current practices, which rely heavily on centralized controls.

EQUITY AND ACCESS TO XR TECHNOLOGY

If developed and distributed correctly, XR has enormous potential to increase accessibility, enable more equal access to virtual experiences, promote inclusivity, and improve user experience. In order to aid the positive benefits, stakeholders need to keep engaging in discussions about international development, education, and diversity, equity, and inclusion, alongside broader conversations about access to underlying technologies (e.g., fifth-generation 5G technology) necessary for inclusive and safe adoption in communities traditionally excluded from early access to novel technologies.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This annex reflects contributions from the following members of the Task Force for a Trustworthy Future Web: Alex Feerst, Murmuration Labs; Camille Francois, Niantic; Sidney Olinsky, Duco Experts; Charlotte Willner, Trust and Safety Professional Association; and Brittan Heller, Digital Forensic Research Lab, as well as the following contributing experts to the task force: Eric Davis, Trust and Safety Professional Association; David Sullivan, Digital Trust and Safety Partnership; Matthew Soeth, Spectrum Labs; and Sara Grimes, University of Toronto. This report includes expert analysis from Duco, whose mission is to empower leading companies to operate safely, securely, and responsibly by mobilizing the world's leading experts to help solve complex challenges.

This report does not represent the individual opinion of any contributor, member of the task force, or contributing organization to the task force. Rather, it serves to consolidate collective research, feedback, and contributions gathered over a five-month period. The contributors are grateful to additional members of the task force and outside experts for their review and feedback.