



SCALING

TRUST

ON

THE

WEB

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



A NOTE FROM THE TASK FORCE DIRECTOR

Digital technologies continue to evolve at breakneck speed, unleashing a dizzying array of society-wide impacts in their wake. In the last quarter of 2022 alone: Meta, Accenture, and Microsoft announced a massive partnership to establish immersive spaces for enterprise environments; Elon Musk took over Twitter; the third-largest cryptocurrency exchange in the world collapsed overnight; the European Union’s landmark Digital Services Act came into force; and generative artificial intelligence (“GAI”) tools were released to the public for the first time. Within a fifty-day span, the outline of a new internet age came into sharper focus.

In December 2022, the Atlantic Council’s Digital Forensic Research Lab began to assemble a diverse array of experts who could generate an action-oriented agenda for future online spaces that can better protect users’ rights, support innovation, and incorporate trust and safety principles—and do so quickly. **The Task Force for a Trustworthy Future Web launched in February, bringing together more than forty experts in policy, AI, trust and safety, advertising, gaming, civil rights, human rights, law, virtual reality, children’s rights, encryption, information security, community organizing, product design, digital currency, Web3, national security, philanthropy, foreign assistance, and foreign affairs.**

Over a five-month sprint, through interviews, expert roundtables, thematic discussions, document reviews, and briefings, task force members shared hard won lessons about what has worked and what hasn’t worked over twenty years of striving to build safe, useful spaces where humans can come together online. **This sprint had four goals:**

- 1 Map systems-level dynamics and gaps that will continue to impact the trustworthiness and usefulness of online spaces regardless of technological change.
- 2 Highlight where existing approaches will not adequately meet future needs, particularly given the emergence of new “metaversal” and GAI technologies and the diversification of online spaces.
- 3 Identify significant points of consensus across the membership’s broad range of perspectives and expertise.
- 4 Generate concrete recommendations for immediate interventions that could fill systems-level gaps and catalyze safer, more trustworthy online spaces, now and in the future.

The task force specifically considered the emerging field of “trust and safety” (T&S) and how it can be leveraged moving forward. That field provides deep insights into the complex dynamics that have underpinned building, maintaining, and growing online spaces to date. Moreover, the work of T&S practitioners, in concert with civil society and other counterparts, now rests at the heart of transformative new regulatory models that will help define how technology is developed in the twenty-first century.

This executive report captures the task force’s key findings and provides a short overview of the truths, trends, risks, and opportunities that task force members believe will influence the building of online spaces in the immediate, near, and medium term. It also summarizes the task force’s recommendations for specific, actionable interventions that could help to overcome systems gaps the task force identified. Given the many ongoing initiatives aimed at developing broad principles, standards, frameworks, or best practices, the task force chose instead to focus primarily on recommendations where philanthropic investment could play an immediate and catalytic role. This executive report provides the introduction to *Scaling*

Trust on the Web, the comprehensive report produced by the task force, which includes six annexes highlighting issues that received special focus:

- 1 A review of how the current T&S field has emerged, the knowledge and practices that have been developed within it, and where it offers opportunity as well as requires evolution and advancement.
- 2 An analysis of where tooling necessary for T&S might benefit from intentional and collective investment and focus.
- 3 An examination of the role that children’s rights and inclusionary participation models can play in debates regarding child safety online.
- 4 An introduction to the gaming industry, highlighting its influence on online spaces now and in the future.
- 5 An assessment of the T&S capabilities of federated platforms, with a particular focus on their ability to address risks like coordinated manipulation and disinformation.
- 6 A review of lessons that could be learned from the evolution of the cybersecurity industry, as well as a forecast of how generative AI may impact T&S.

I am indebted to the task force’s members, contributing expert organizations, and contributing experts for their time, care, candor, creativity, wisdom, and overall esprit de corps throughout this fast-paced and iterative endeavor. I am also deeply grateful to Nikta Khani, associate director of the task force, as well as to Rose Jackson, Eric Baker, and Graham Brookie of Digital Forensic Research Lab, and Mary Kate Alyward of the Atlantic Council, for their superlative support, guidance, and diligence.

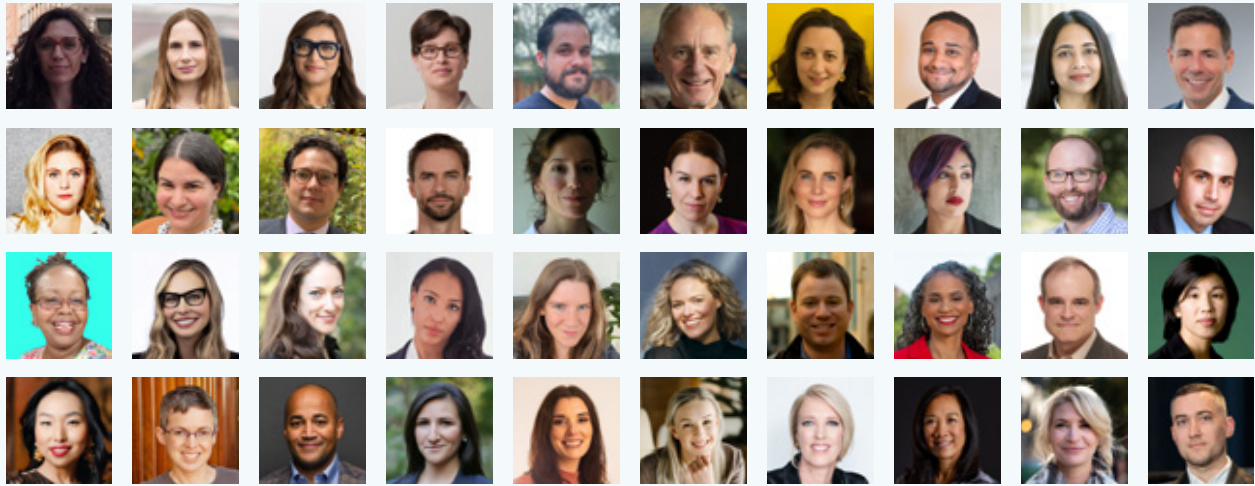
By looking beyond any particular challenge to the incentive structures defining—and constraining—the construction of our collective digital future, this task force has clarified where gaps in understanding or incentives must be addressed to further important change in building safer online spaces. Critically, the task force took on as a baseline assumption the inherent dignity and importance of stakeholders whose rights and perspectives have historically been ignored in the creation of existing online spaces—key among them marginalized communities in the Global North, entire populations in the “[Global Majority](#),” women, and youth.

Naming a problem makes it easier to solve. Clarifying a challenge makes it easier to overcome. Identifying an opportunity makes it easier to realize. This task force has named problems; clarified challenges; and identified opportunities. It is my greatest hope that the findings presented in *Scaling Trust on the Web* spur renewed and refreshed dialogue, collaboration, and innovation, as well as material investments in realizing the task force’s key recommendations.

Kat Duffy
 Director
 Task Force for a Trustworthy Future Web

TASK FORCE MEMBERS

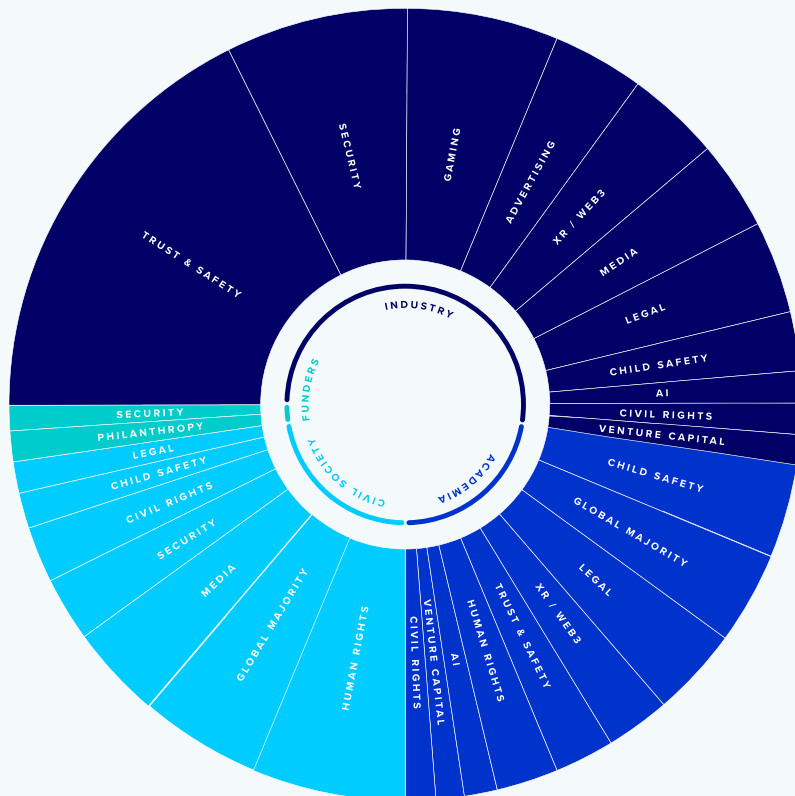
The task force comprises forty experts across industry, civil society, academia, and philanthropy. Every task force member brings deep expertise in at least two (and often three or four) of the following areas: policy development, AI, trust and safety, advertising, gaming, civil rights, human rights, law, virtual reality, children's rights, encryption, information security, community organizing, product design, digital currency, Web3, national security, philanthropy, foreign assistance, and foreign affairs. Task force members were chosen not only for having subject matter expertise, but also for bringing seasoned, nuanced perspectives to profoundly complex challenges. The task force's findings were enriched by the input of fifteen contributing expert organizations as well as dozens of additional contributing experts.



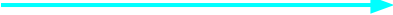
EXPERTISE OF THE TASK FORCE MEMBERS

SECTOR REPRESENTED

- INDUSTRY
- ACADEMIA
- CIVIL SOCIETY
- FUNDERS



EXECUTIVE SUMMARY



That which occurs offline will occur online. Now and in the future, some online spaces will inevitably evolve into arenas hosting a fierce contestation of norms. Moreover, in any democratic society, online or offline, **some harms and risks simply must be accepted as a key principle of protecting the fundamental freedoms that underpin that society.** No technology has solved long-standing and deeply-rooted societal problems such as racism, sexism, ethnic hatred, intolerance, bigotry, or struggles for power. No technology is likely to do so in the future.

It is equally true that **choices made when creating or maintaining online spaces generate risks, harms, and beneficial impacts.** These choices may rest in policy determinations, product designs, operational systems, organizational values, revenue models, or other strategic decisions. **These choices are not value neutral, because the resulting products, platforms, and technologies do not enter into neutral societies. Malignancy migrates, and harms are not equally distributed across societies.** Marginalized communities (however they might be constituted in any particular country or culture) suffer disproportionate levels of harm online and off. Online spaces that do not acknowledge or plan for that reality consequently scale malignancy and marginalization.

We are at a pivotal moment in the evolution of online spaces.

From the dramatic expansion of access to generative AI tools in only six months since ChatGPT was released publicly, to the increased popularity of decentralized platforms such as Mastodon or Bluesky, to the coming normalization of immersive environments for social and professional gathering, the speed and scale of change are increasing exponentially. Major regulation from the European Union (EU) and other key jurisdictions is creating new incentives and driving new practices across the technology industry, and yet, no consensus exists on what “good” should look like in the digital world of today, let alone in the future. Moreover, governmental action has historically proven incapable of keeping pace with emerging technology (unless that action has been to censor, surveil, block, or otherwise violate fundamental rights and freedoms).

Risk and harm are set to increase at an exponential pace, and existing institutions, systems, and market drivers cannot keep up. Industry will continue to drive these rapid changes, but is likely to be unable or unwilling to solve the core problems at hand. In response, innovations in governance, research, financial, and inclusion models must scale with similar velocity.

Thankfully, **the knowledge needed to identify and build solutions has been developing steadily both inside companies and outside of them.** Significant collective expertise now exists to illuminate not

only where harms and risks can scale through existing and emerging technologies, but also where lessons learned can be applied proactively to construct safer, more trustworthy spaces. Within industry, “trust and safety” (T&S) practitioners with deep insight into the complexities of building and operating online spaces are rapidly evolving from an insular community into a professional and newly accessible field. Outside industry, civil society groups, independent researchers, and academics continue to lead the way in building collective understanding of how risks propagate via platforms—and how products could be constructed to better promote social well-being and to mitigate harms—especially within marginalized communities.

These statements represent some of the greatest points of consensus across the Digital Forensic Research Lab’s Task Force for a Trustworthy Future Web, which brought together more than forty experts in technology policy, artificial intelligence (AI), trust and safety, advertising, gaming, civil rights, human rights, law, virtual reality (VR), children’s rights, encryption, information security, community organizing, product design, digital currency, Web3, national security, philanthropy, foreign assistance, and foreign affairs.

From January to May of 2023, the task force conducted a sprint to accomplish four goals:

- ▶ Map systems-level dynamics and gaps that will impact the trustworthiness and usefulness of online spaces regardless of technological change.
- ▶ Highlight where existing approaches will not adequately meet future needs, particularly given the emergence of new “metaversal” and generative AI (GAI) technologies and the diversification of online spaces.
- ▶ Identify significant points of consensus across the membership’s broad range of perspectives.
- ▶ Generate concrete recommendations for immediate interventions that could catalyze safer, more trustworthy online spaces, now and in the future.

Task force members were joined by representatives from fifteen contributing organizations as well as dozens of contributing experts, who participated in interviews, expert roundtables, thematic discussions, document reviews, and briefings. ***Scaling Trust on the Web*, the task force’s comprehensive report, captures the results of that exercise, and reflects hard-won lessons from more than twenty years of building spaces where humans come together online.** Those include the following, additional key findings:

- 1 An emerging T&S field creates important new opportunities for collaboration.
- 2 Academia, media, and civil society bring crucial expertise to building better online spaces.
- 3 Protecting healthy online spaces requires protecting the individuals who defend them.
- 4 Learning from mature, adjacent fields will accelerate progress.
- 5 The gaming industry offers unique potential for insights and innovation.
- 6 Existing harms will evolve and new harms will arise as technologies advance.
- 7 Systemic harm is exacerbated by market failures that must be addressed.
- 8 Philanthropies and governments can shape incentives and fill gaps.

Acknowledging that the philanthropic sector is uniquely capable of catalyzing novel and creative pathways to supporting systems-level change, the task force also recommended significant and immediate investments designed to:

- 1 Craft and implement initiatives that target market failures and incentives gaps.
- 2 Accelerate the maturation/professionalization of trust and safety as an independent field.

- 3 Break down knowledge silos and share information and expertise.
- 4 Protect and grow the enabling environment necessary to innovate more trustworthy, useful online spaces.
- 5 Expand investment in proactive, future-facing research and initiatives.

We are on the precipice of a new digital era. It is our hope that the insights captured in *Scaling Trust on the Web* galvanize investments in systems-level solutions that reflect the expanding communities dedicated to protecting trust and safety on the web, the trailblazers envisioning the next frontier of digital tools and systems, and the rights holders whose futures are at stake.

TABLE OF CONTENTS

A Note from the Task Force Director	1
Executive Summary	4
Table of Contents	7
Annex Directory	8
Introduction	9
KEY FINDING 1 The Emergence of a Trust and Safety Field Creates Important Opportunity	13
New Initiatives Are Shifting T&S from a Community of Practice into a Field	13
Accelerating Knowledge Sharing About T&S Practices Is a Critical Need	14
Building Openly Available Tooling Is an Area of Opportunity	15
T&S Would Benefit from a Deeper and More Diverse Professional Pipeline	16
KEY FINDING 2 Academia, Media, and Civil Society Bring Crucial Expertise to Building Better Online Spaces	17
Academic Researchers Need Improved Access to Support T&S	17
Civil Society Expertise Is Crucial and Under Threat	18
The Media Is Fundamental to Improving Understanding and Accountability, and Also Under Threat	19
KEY FINDING 3 Protecting Healthy Online Spaces Requires Protecting the Individuals who Defend Them	21
KEY FINDING 4 Learning from Mature, Adjacent Fields Will Accelerate Progress	23
Cybersecurity	23
Human Rights	24
Additional Key Fields	25
KEY FINDING 5 The Gaming Industry Offers Unique Potential for Insights and Innovation	27
KEY FINDING 6 Existing Harms Will Evolve and New Harms Will Arise as Technologies Advance	29
Federated Spaces	29
Immersive Spaces	31
Generative AI	32
KEY FINDING 7 Systemic Harm Is Driven by Market Failures That Must Be Addressed	34
Measuring T&S Is a Meaningful Challenge	34
Emerging Regulation Is Already a Market Driver for T&S	35
The Role of Venture Capital Has Been Underexamined	37
KEY FINDING 8 Philanthropies and Governments Can Shape Incentives and Fill Gaps	38
Key Recommendations	39
Conclusion	46
Acknowledgments	47
Task Force Members	48
Expert Contributing Organizations	49
Contributing Experts	49

ANNEX DIRECTORY

Annex 1

CURRENT STATE OF TRUST AND SAFETY

Annex 2

BUILDING OPEN TRUST AND SAFETY TOOLS

Annex 3

RESPECTING CHILDREN AS RIGHTS HOLDERS

Annex 4

DECONSTRUCTING THE GAMING ECOSYSTEM

Annex 5

COLLECTIVE SECURITY IN A FEDERATED WORLD

Annex 6

**LEARNING FROM CYBERSECURITY,
PREPARING FOR GENERATIVE AI**

INTRODUCTION

WE HAVE A NARROW WINDOW AND OPPORTUNITY TO LEVERAGE DECADES OF HARD WON LESSONS AND INVEST IN REINFORCING HUMAN DIGNITY AND SOCIETAL RESILIENCE GLOBALLY.

→ The digital future—and any trustworthy future web—will reflect all of the complexity and impossibility that would be inherent in understanding and building a trustworthy world offline. No technology has solved long-standing and deeply rooted societal problems such as racism, sexism, ethnic hatred, intolerance, bigotry, or struggles for power. No technology is likely to do so in the future.

This hard truth represents one of the greatest areas of consensus within the Task Force for a Trustworthy Future Web: that which occurs offline will occur online. Now and in the future, some online spaces will inevitably evolve into arenas hosting a fierce contestation of norms. Moreover, in any democratic society, online or offline, some harms and risks must be accepted as a key principle of protecting the fundamental freedoms that underpin that society.

This leads to a second, equally significant area of consensus: it is also true that choices made when building or maintaining online spaces play a critical role in accelerating or mitigating risks, harms, and beneficial impacts. Existing and future online spaces must be better at protecting users' rights, supporting innovation, and incorporating trust and safety¹ (T&S) principles—and do so quickly. Policy determinations, product designs, operational systems, organizational values, revenue models—these choices are not value neutral because the products that result do not enter neutral societies. **Malignancy migrates. Harms are not equally distributed across societies. Marginalized communities (however they might be constituted in any particular country or culture) suffer**

¹ Please see below for more on T&S, as well as [Annex 1: Current State of Trust and Safety](#).

disproportionate levels of harm online and off. Online spaces that do not acknowledge or plan for that reality consequently scale malignancy and marginalization by design.

This inspired a third major area of consensus: **risk and harm are currently set to scale and accelerate at an exponential pace, and existing institutions, systems, and market drivers cannot keep pace.** Industry will likely continue to drive these rapid changes, but also prove unable or unwilling to solve the core problems at hand. Innovations in governance, research, financial, and inclusion models must scale with similar velocity. By developing more creative and aggressive strategies, philanthropies and governments can play a significant role in meeting this moment more effectively.

A NOTE ON SCOPE AND TERMINOLOGY

A trustworthy future web will encompass a far wider range of technologies than the task force could reasonably cover. The task force limited its scope to considering internet-based spaces now and in the future that bring people together.

Although “platform” is arguably the most widely used term to describe spaces online where people come together, the term’s close connection to social media does not serve the broader goals of the task force’s inquiry or this report. Consequently, “online spaces” and “platforms” will be used interchangeably to signal the wide range of possibilities that exist beyond traditional social media.

This report frequently uses “companies” to refer to the organizations or entities that control an online space. It is worth noting that while most platforms are run by corporate entities, notable exceptions exist, such as the nonprofit Wikimedia Foundation.

“[Global Majority](#)” is used throughout this report rather than terms such as “Global South,” “Developing World,” or the particularly egregious phrase common to the tech industry, “Rest of World.” More information about the origins and meaning of the term can be found [here](#). For the purposes of this report, “Global Majority” refers to the vast majority of the world’s population who do not come from majority White, wealthy nations or regions, as well as to individuals and communities who are marginalized within those nations/regions. “Global North” is used to reference those nations/regions.

FOCUS AREAS OF TASK FORCE

WEARABLES

SMART DEVICES

CRYPTOCURRENCY

SHARING ECONOMY PLATFORMS

CONSUMER-FOCUSED MESSAGING

FEDERATED SPACES

GAMING

XR/IMMERSIVE SPACES

SOCIAL MEDIA PLATFORMS

E-CONVENING PLATFORMS

DATING APPS

APP STORES

SEARCH ENGINES

E-COMMERCE PLATFORMS

CLOUD SERVICE PROVIDERS

- PRIMARY FOCUS
- SECONDARY FOCUS
- OUT OF SCOPE

Across the task force, there was a strong consensus that we are at a pivotal moment in the evolution of online spaces. Major regulation from the European Union and elsewhere is creating new incentives and driving new practices across the technology industry. At the time of this publishing, companies are reallocating resources, teams, and approaches across T&S matters because of these new rules. And yet, no consensus exists on what “good” should look like in the digital world of today, let alone in the future.

Governmental action has perennially proven incapable of keeping pace with emerging technology (unless that action has been to censor, surveil, block, or otherwise violate fundamental rights and freedoms). While there are some established answers to known challenges, newer, faster, and more challenging questions continue to emerge for industry, civil society, and government to answer. From the dramatic expansion of access to GAI tools in only the six months since ChatGPT was released publicly, to the increased popularity of decentralized platforms such as Mastodon or Bluesky in the six months since Twitter’s dismantling of T&S teams and processes, to the coming normalization of immersive environments for social and professional gathering, the speed and scale of change are increasing exponentially.

Thankfully, **the knowledge needed to identify solutions has been developing steadily inside the technology industry and outside of it**, evolving into a diverse ecosystem with the expertise to illuminate not only where harms and risks can scale through existing and emerging technologies, but also where lessons learned can be applied proactively to construct safer, more trustworthy spaces. A community of T&S practitioners, who can offer deep insight into the complexities of building and operating online spaces within industry, is steadily evolving into a professional field. As this field emerges, it is creating new potential within a broader ecosystem of experts to expedite transformative collaborations, knowledge sharing, and innovation.²

This rare combination of regulatory sea change that will transform markets, landmarks in technological development, and newly consolidating expertise can open a window into a new and better future, in which the next wave of connective technology brings innovation and systemic resilience into better balance. **It is within this context that the task force arrived at the following key findings:**³

- 1 The emergence of a trust and safety field creates important opportunity.
- 2 Academia, media, and civil society bring crucial expertise to building better online spaces.
- 3 Protecting healthy online spaces requires protecting the individuals who defend them.
- 4 Learning from mature, adjacent fields will accelerate progress.
- 5 The gaming industry offers unique potential for insights and innovation.
- 6 Existing harms will evolve and new harms will arise as technologies advance.
- 7 Systemic harm is driven by market failures that must be addressed.
- 8 Philanthropies and governments can shape incentives and fill gaps.

The task force also developed a series of concrete recommendations given the urgent need for action across a wide constellation of sectors and fields. The task force focused particularly on recommendations for philanthropic investments to fill systemic gaps because the philanthropic sector is uniquely capable of

² Due to the insights the emerging T&S field can provide into the complex dynamics that have underpinned building, maintaining, and growing online spaces to date, the task force specifically considered the emerging field of T&S and how it can be leveraged moving forward. This report consequently relies heavily on T&S as a framing mechanism. That *does not* reflect a consensus across the task force that T&S alone provides an adequate frame for future web design, nor does it reflect a consensus that T&S is superior to alternative lenses of inquiry—such as technology and democracy, technology and human rights, a feminist internet, decolonization, ethical tech, responsible tech, or other noteworthy constructs. These alternative framings play a valid and important role in building a vision for a more equitable digital future. This report’s focus on T&S is not meant to take away from their legitimacy or importance.

³ Findings are ordered to facilitate narrative flow. They do not reflect any hierarchical structure.

catalyzing novel and creative pathways to achieving systems-level change. The task force urged significant and immediate investments designed to:

- 1 Craft and implement initiatives that target market failures and incentives gaps.
- 2 Accelerate the maturation/professionalization of trust and safety as an independent field.
- 3 Break down knowledge silos and share information and expertise.
- 4 Protect and grow the enabling environment necessary to innovate more trustworthy, useful online spaces.
- 5 Expand investment in proactive, future-facing research and initiatives.

WHAT IS “TRUST AND SAFETY”?

For decades, an area of specialty and practice⁴ that is increasingly referred to as “Trust & Safety” (T&S) has developed inside US technology companies to diagnose and address the risks and harms that face individuals, companies, and now—increasingly—societies on any particular online platform.

No single definition of T&S holds across all audiences. Stated most generally, T&S anticipates, manages, and mitigates the risks and harms that may occur through using a platform, whereas “cybersecurity” and “information security” address attacks from an external actor against a platform.

A T&S construct may describe a range of different verticals or approaches. “Ethical” or “responsible” tech; information integrity; user safety; brand safety; privacy engineering—all of these could fall within a T&S umbrella. T&S practice is equally varied and can include a variety of cross-disciplinary elements ranging from defining policies, to rules enforcement and appeals, to law enforcement responses, community management, or product support.

The types of harms that T&S may take on (when considering online spaces) include coordinated inauthentic behavior, copyright infringement, counterfeiting, cross-platform abuse, child sexual abuse material (CSAM), denials of service (DOS) / distributed denials of service (DDOS), disinformation, doxing, fraud, gender-based violence, glorification of violence, harassment, hate speech, impersonation, incitement to violence or violent sentiment, misinformation, nonconsensual intimate imagery, spam, synthetic media (for example, deepfakes), trolling, terrorist and violent extremist content (TVEC), violent threats, and more. These harms are specific to online spaces and are not meant to denote the range of harms that T&S considers as a field.

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 American value sets. This American 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 the United States 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.

⁴ See *Annex 1: Current State of Trust and Safety*, for a more comprehensive overview of this field, including the origin of the term “Trust and Safety.” For excellent overviews of the evolution of trust and safety, see “Introducing the Trust and Safety Curriculum,” Trust and Safety Professional Association, June 17, 2021; “Knowledge Hub: Trust & Safety,” All Tech Is Human, n.d.; Data & Society’s *Origins of Trust and Safety* (podcast), No. 134 (2020); Kate Klonick, “The End of the Golden Age of Tech Accountability, The Klonickles (newsletter), March 3, 2023; and “The Trust and Safety Teaching Consortium,” Stanford Internet Observatory, n.d.

KEY FINDING 1

THE EMERGENCE OF A TRUST AND SAFETY FIELD CREATES IMPORTANT OPPORTUNITY

Because US technology companies were at the forefront of building and scaling online spaces, that industry was the first to achieve massive scale for users and revenue. By extension, that same industry also had unique capacity to propagate harm and to innovate ways to mitigate harm, and the earliest exposure to external scrutiny, regulatory pressures, and business risks. That is why, over decades, a community of practice has developed within US technology companies to identify and address the risks and harms that face individuals, companies, and now increasingly societies on any particular online platform. Generally referred to as trust and safety (T&S), this emerging field has also served as a sandbox for piloting and refining a range of policies, products, tools, and mechanisms aimed at constructing online spaces that can better promote social well-being and mitigate harmful content, behavior, and other externalities.

Commitments to T&S are increasingly seen as an organizational baseline for the responsible running of a platform. The emerging field of T&S can and should be leveraged to help construct online platforms and digital technologies that better promote social well-being and that mitigate harmful content, behavior, and other externalities, in particular harms impacting marginalized communities. **For more than a decade, T&S expertise has been trapped largely within niche communities of practice inside large companies. As the community of practice is expanding and evolving into a professional field, that knowledge is finally seeing the light of day, and creating new opportunities for action and collaboration.**

NEW INITIATIVES ARE SHIFTING T&S FROM A COMMUNITY OF PRACTICE INTO A FIELD

While T&S has essentially existed as long as internet services have, it operated for many years as an insular, if growing, community of practice. In recent years, new initiatives have begun to shape that community into an emerging professional field. The number of organizations, courses, and initiatives supporting the evolution and development of T&S has been expanding dramatically and consistently over the past several years. The [Trust and Safety Professional Association](#) and its

concomitant foundation launched in 2020 to support the global community of T&S professionals and to improve “society’s understanding of T&S,” respectively. Spectrum Labs’ [#TSCollective](#) has [emerged](#) as a community of more than 700 T&S professionals, dedicated to supporting knowledge sharing as well as community building within T&S.

Former Facebook integrity team and product team workers [launched](#) the [Integrity Institute](#) with the goal of bringing together industry professionals to “advance the theory and practice of protecting the social internet,” and the [Oasis Consortium](#) formed and developed [standards](#) that companies could use to support user safety across online spaces. Leading US technology companies also formed the industry-based [Digital Trust and Safety Partnership](#), which has since launched a [T&S assessment framework](#), an [inaugural evaluation of T&S best practices](#), and a [T&S glossary of terms](#) that will be finalized in 2023.

At least four new T&S conferences also launched in 2022: the inaugural [TrustCon](#), the [Trust and Safety Research Conference](#) in the United States, the [Safety Matter Summit](#) (now called the ProSocial Summit), and the [Trust and Safety Forum](#) in Europe. Within academia, the Stanford Internet Observatory created the [Journal of Online Trust and Safety](#). Stanford University also launched the first undergraduate course in [Trust and Safety Engineering](#), and a new [Open Source T&S course](#); Columbia University began offering a [graduate level T&S course](#), [New York University a T&S certificate program in collaboration with ActiveFence](#); and Griffith College in Ireland a [postgraduate diploma program](#). In addition, podcasts, substacks, blogs, hackathons, and a range of other endeavors (including a popular content moderation [game](#)) have continued to emerge from the T&S community.

These new formal and informal structures open T&S practice up to a wider array of stakeholders. **New channels for information exchange and learning exist in 2023 that can be transformative not only within the T&S practitioner community, but also between T&S and a wider community of experts in civil society, media, academia, and the public sector who share similar goals for online spaces.**

ACCELERATING KNOWLEDGE SHARING ABOUT T&S PRACTICES IS A CRITICAL NEED

Organizations that create intentional space for T&S practitioners to learn from each other and build community play a meaningful role in moving T&S forward as a field. **Historically, practitioners have had to rely primarily on informal (often opaque) exchanges within their networks as a primary means of learning best practices for a wide range of topics.** This includes policy development, product design, T&S tooling, regulatory compliance, and external engagement. It extends, though, to broader business practices, such as improving knowledge around structuring T&S within an organization; where in a company’s scale or maturity model it should expect new T&S challenges to arise; and where early strategic investments in T&S are most effective and most critical. Having access to a more formalized body of knowledge and opportunities for community engagement is particularly important for practitioners who move from large companies to smaller companies or start-ups, and consequently have less access to in-house institutional knowledge or other resources.

Given the current rise in regulatory requirements, audits, and assessments will increasingly inform T&S practices within companies. From [overarching assessment frameworks to transparency, due diligence, user safety, or human rights impact assessments](#) (among other [possibilities](#)), this move toward more standardized approaches and focus will move T&S toward greater coherence in ways that can aid information sharing among practitioners and between different stakeholder groups. Companies are already investing in processes and structures that will help ensure regulatory compliance, as well as the capacity to respond to audit or assessment findings; the need for a rapid escalation in expertise will be significant.

Auditors, assessors, vendors, and advisers will represent a growing segment of the broader T&S services industry in the coming years. This creates a very real risk that influence will consolidate even further with-

in industry and its direct affiliates (i.e., auditing companies) in the Global North. Global Majority-based experts must be empowered to develop frameworks and assessments that proactively measure risks, harms, and opportunities that would otherwise be invisible to T&S teams, auditors, and assessors. This will play a meaningful role in overcoming long-standing, at times catastrophic, power imbalances between the companies building online spaces and the communities impacted by them.

Finally, the role government has played in shaping T&S merits deeper and more consistent, transparent analysis. Much of T&S evolution to date has been defined by the complex responses that platforms must design in the face of governmental requests for content takedowns, user data, or abuses of the platform by state actors. Governments have demanded platforms' compliance with laws or regulations that violate human rights, or with the laws of a country where a company is headquartered. Independent researchers, civil society activists, and T&S practitioners across the task force emphasized that as governments push for greater transparency from companies, they must also demonstrate leadership in ensuring that their own policies, priorities, and practices when engaging with online spaces reflect a greater accountability to the citizens they represent, and to their citizens' fundamental human rights—as well as providing greater backing to platforms when platform users' fundamental rights are under attack. Supporting collaborations that further greater knowledge sharing on this point would help further systems-level responses rather than laying this burden solely at the feet of individual companies and their T&S teams.

BUILDING OPENLY AVAILABLE TOOLING IS AN AREA OF OPPORTUNITY

T&S requires a technical implementation layer that can become highly complex quite quickly, and is often built out with homegrown tooling suites and organizational structures over time 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 iterative looping through four distinct goals: detection, enforcement, measurement, and transparency (i.e., documentation/communication).⁵

The logistical aspects of T&S operations could benefit from the development of robust open tooling.⁶ Providing access to a suite of basic but useful tools would be of significant benefit to small- and medium-sized companies that may want to build a strong foundation for eventual T&S teams and tools, but lack the resources to invest early in solving for problems that will occur at a later stage of growth. For example, hash-matching tools that could detect exact and near-exact matches of previously identified content, or tool kits that could help build classifiers to assess new, not previously seen content or behavior, could also be of widespread benefit. Finally, building tools that could allow external experts, such as researchers, to provide information to multiple platforms through one pipeline would greatly improve efficiencies in the broader ecosystem. This could be particularly powerful for civil society and academic researchers tracking abusive actors across platforms.

More effective, openly available tooling—as well as more accessible guidance on best practices for development of T&S tools—could lower barriers to the development of, and increase competition among, a diversity of services. This could meaningfully change the degree to which each organization must reinvent the wheel for in-house solutions. **It could also help address what is essentially a market failure: individual services may not internalize all the social costs of harms occurring on their platform, and thus may not invest sufficiently in socially optimal T&S.**

⁵ T&S tooling can also be thought of in terms of a “tech stack.” A tech stack is a set of tools that serves particular purposes and is aligned to a product development process, which can broadly be generalized to back-end, mid-layer, and front-end components. See e.g., [Zoom's discussion of its T&S “tech stack.”](#) From this vantage point, detection, confirmation and enforcement, measurement, and transparency are the relevant goals of the “stack.”

⁶ For a deep dive into this topic, please see [Annex 2: Open Tooling.](#)

There are limitations to what can be supported through openly available tooling. In particular, content-specific detection tools present a complex challenge, especially with regard to overall governance and institutional support. While a wide array of services may have policies against common types of content (e.g., hate speech), services' individual policies vary and no one tool will suit all. Detection tools must be updated consistently over time. **Task force members emphasized that “set it and forget” is an impossibility within T&S tooling and practice.** Moreover, these tools may raise complex legal questions—for instance, how to balance the privacy implications of processing personal data. In turn, creating shared databases of violative content or content-specific classifiers raises many questions beyond simply technological design. While this is a more complex endeavor, it can provide significant utility.

T&S WOULD BENEFIT FROM A DEEPER AND MORE DIVERSE PROFESSIONAL PIPELINE

As with many fields, **a more robust and diverse talent pipeline is urgently needed to support the expansion of T&S practices and principles across a broader array of teams, products, and research initiatives.** Given its long-standing American cultural roots, T&S would benefit from building greater geographic diversity into HQ-based teams. Frontline content moderation workers (described in more detail below) also bring powerful expertise into the T&S space because of the vast range of cultures, languages, and communities they represent. Diverse perspectives play a crucial role in identifying emerging threats, differentiating harms, clarifying contextual questions (e.g., is a trending hashtag hate speech or cultural reappropriation?), and crafting proportionate responses that reflect a particular platform's policies.

The next generation of T&S practitioners and experts should also come from a more diverse range of disciplines. This will help T&S respond to the diversity of challenges present in AI and metaversal technologies (such as decentralized and/or immersive environments), as well as the increasingly varied range of societal harms online platforms can exacerbate. The creation of new university programs at the undergraduate and graduate levels will be critical in increasing the breadth of technical, geopolitical, and cultural expertise necessary for the field to flourish in the future. It is important that such programs not remain limited to elite institutions in the United States and Western Europe, but rather extend to venues such as community colleges, as well as to educational institutions across other global regions. Geographic diversity will support more contextualized research and enable a wider range of students and scholars to inform the field's development. Supporting the inclusion of more experts in elections, journalism, human rights, health, and other key societal sectors will also be key for the T&S field.

Task force members emphasized that **moves to formalize and professionalize T&S could create barriers to entry and cement elitism into an emerging field that will rely on diverse perspectives in order to mature effectively. These dynamics and trade-offs should be taken into consideration when considering formalized growth.**

KEY FINDING 2

ACADEMIA, MEDIA, AND CIVIL SOCIETY BRING CRUCIAL EXPERTISE TO BUILDING BETTER ONLINE SPACES

→ 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 (nontechnological) side of a problem.

T&S is only one component of a much broader universe of actors and experts who have also played a critical role in identifying and mitigating harm, including activists, researchers, academics, and journalists.

These sectors bring crucial expertise into addressing challenges such as hate speech, harassment, and defamation; mis- and disinformation; child sexual abuse material and nonconsensual intimate imagery; terrorist or violent content; or trolling, brigading, and impersonation, among others.⁷ These stakeholders, among them policymakers, researchers, and civil society advocates, may rely on frameworks such as “platform accountability,” “platform governance,” “responsible tech,” and “ethical tech,” to articulate the concerns that most companies would address through a T&S lens. Any vision for a future with safer, more trustworthy online spaces must include a clear vision for recognizing the insights and influence of this broader community of experts.

ACADEMIC RESEARCHERS NEED IMPROVED ACCESS TO SUPPORT T&S

The budding T&S academic initiatives described above (e.g., courses, journals, research conferences) are essential at a moment when the

⁷ For a quick review of the common types of abuse, enforcement practices, and key practices within T&S today, please see the final page of this annex as well as the Digital Trust & Safety Partnership’s public consultation [Glossary of Trust and Safety Terms](#).

gap between practitioners and the academic community is large. Projects and conversations to help close this gap have in the past focused on access to data for researchers, and on specific subareas of T&S seen as deserving of immediate and enhanced accountability (e.g., disinformation). This helps, **but more must be done to help ensure that practitioners are better informed by academic research relevant to their fields and, in turn, ensure that academic research can be shaped by an accurate understanding of the broader systems used across T&S functions.** As highlighted above in the section on diversifying T&S pipelines, work and investments in this area should not be limited to elite, Global North institutions, but should instead help deepen academic research capacity and independence across educational institutions in the Global Majority countries. Entirely new areas of research/specialization also cry out for development, such as prosocial design, human computer interaction, online measurement, and forensics based on open source intelligence.

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. With regulation such as the EU's [Digital Services Act \(DSA\)](#) calling for more transparency and access to data around moderation practices, **it is imperative to invent new systems that will support transparent access to the broader information (not just outcomes data) needed for researchers to help innovation and accountability across the different subareas of T&S.**

CIVIL SOCIETY EXPERTISE IS CRUCIAL AND UNDER THREAT

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,⁸ 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 also have 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 functions as a major lever for actioning 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](#) have helped drive forward debate and consensus around best practices and minimum acceptable standards for companies. Nongovernmental organizations 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 safety](#) 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—remain a significant and counterproductive gap within industry.** The onus continuously rests on civil society—which as a field comprises organizations that are generally smaller, less-well resourced, and 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 devel-

⁸ For more on the role of civil society, please see [Annex 1: Current State of Trust and Safety](#).

opment or the [reporting of “bad” content](#) have been the most common mechanisms that companies have piloted, but they have not proven to be 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, as complementing the capacities of professional T&S teams that rely on them for analysis and to understand issues like societal-level risks or specific bad actors. Companies’ internal systems are often not tailored for the needs of partners from the Majority World, and not enough has been done to engage such partners proactively in anticipating the evolution of local risk factors, harms, and user needs. For example, companies whose primary revenue-driving markets are English-language and culturally Western have proven unlikely to invest in building high-quality classifiers for other markets and languages, rendering the efficacy and nuance of such tools less valuable. Collaborations with civil society to solve for this problem could bring new approaches to light. Civil society can also play a particularly important role in identifying how harms operate and evolve across platforms—an analysis that T&S teams inside of 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. A healthy digital future depends on such independent and contextualized knowledge. And yet, civic space is [under attack](#) globally, degrading the capacity of civil society to operate, let alone participate meaningfully, in [developing](#) trusted and safe spaces online. As [autocracy rises](#) globally, the number of countries where civil society can legally operate is shrinking. Since 2015, approximately one hundred laws have been proposed by governments [targeting](#) the ability of civil society organizations to register, operate, receive foreign funding, or assemble freely. Absent dramatic interventions by companies and donors to ensure civil society support, [funding](#), and engagement, this key sector’s expertise and influence will be increasingly difficult to access.

THE MEDIA IS FUNDAMENTAL TO IMPROVING UNDERSTANDING AND ACCOUNTABILITY, AND ALSO UNDER THREAT

Journalism has been a key stakeholder⁹ in driving attention to T&S, notably in the areas of platform vulnerabilities. There are, however, limitations and shortfalls within the current practice of technology journalism, 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. A classic example of this in the technology industry are the reports about [YouTube and radicalization](#): a slew of media stories connected YouTube’s algorithmic video recommendations to a rise in violent extremism. Despite subsequent research [debunking](#) this relationship, this connection remains a misconception in the general public. More recently, coverage of [AI large language models](#) (LLMs) has led to widespread misunderstandings among nontechnical readers about LLMs’ relationship to human intelligence and emotions. The blame for this lies in part with honest misapprehension and intellectual reckoning with novel technologies; it also lies with lazy regurgitation of sensationalist clickbait.

⁹ For more on the role of media, please see [Annex 1: Current State of Trust and Safety](#). Although not the focus of this section, it should be noted that media outlets have also developed innovative products and tooling to examine the societal impacts of online spaces, similar to academic researchers and civil society organizations.

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. This is detrimental to both the press and tech platforms, as skilled technology journalists have played an important and constructive role in driving public understanding of company incentives and priorities; wielding corrective influence on industry excesses through rigorous investigative reporting; and helping shift broader media coverage away from sensationalism and toward nuanced and informative analysis. Meanwhile, companies' refusal to engage with the press abandons key opportunities to correct inaccuracies and inform a policymaking audience. **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 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. While there are nascent efforts to expand global coverage of technology and society, particularly in underserved geographies and languages, significant need still exists for immediate, material, and sustained investment in shoring up media. Record numbers of journalists were jailed worldwide in 2022, news deserts are expanding in the United States and abroad, and advertising revenue continues to decline for media worldwide—going instead to technology companies. In an era of increasing global autocracy, where platforms are being used as tools for repression, disinformation, and radicalization, a lack of reporters who can identify or elevate specific harms or risks being propagated at local levels through platforms doesn't only elevate risk for those communities and for platforms. It elevates national security, law enforcement, and intelligence risks as well.

KEY FINDING 3

PROTECTING HEALTHY ONLINE SPACES REQUIRES PROTECTING THE INDIVIDUALS WHO DEFEND THEM

→ T&S practitioners,¹⁰ particularly content moderators, face high risks of developing post-traumatic stress disorder, depression, and other psychosocial harms. T&S practitioners who publicly represent a company's position increasingly face targeted public bullying and harassment, as do company leaders and independent researchers. Multiple task force members agreed that this harassment is aimed directly at influencing behavior, politicizing T&S decisions, dissuading research, and chilling practitioners' speech and personal ability to continue supporting T&S work. It also is designed to disincentivize investments, philanthropic or otherwise, in this sector. It's a very troubling trend that the T&S community will need to grapple with for years to come.

The T&S community visible at conferences and in emerging organizations overwhelmingly reflects individuals who hold T&S positions within industry or affiliated sectors. Of the more than one hundred thousand people who work in trust and safety, the majority are in content moderation roles. These frontline content moderators have been referred to as essential gatekeepers of the internet, assessing millions of pieces of content a day. This vast community works primarily for contracting companies across the United States and in countries such as Ireland, India, the Philippines, and Kenya. These individuals play a critical role in T&S, but contracting structures often fail to leverage these moderators' expertise or ensure fair labor practices and humane working conditions. The increased risks that externally contracted content moderators face have long been documented. These workers can face low pay, poor working conditions, exposure to traumatizing content, and often a sense of extreme powerlessness when they are so removed from decision-makers that their insights and warnings remain untapped or ignored.

Implementing workplace wellness programs to address the needs of those exposed to harmful content is paramount; aside from the stated

¹⁰ For more on this topic, see [Annex 1: Current State of Trust and Safety](#).

health impacts on moderators, platforms may face liability and a decrease in productivity if they do not make long-term investments to protect their employees. Indeed, **burnout and declining mental health not only impact the individuals doing the T&S work but the sustainability and maturation of the field as a whole.** Interventions such as image blurring and moderation tools can help improve experiences for human moderators. Though developments in and applications of AI will shift how humans interact with harmful content in content moderation and T&S work, **there will always need to be humans involved in reviewing some content, setting policy, reviewing process, and confirming decisions.** In addition, current models for front-line workers are likely to be replicated for handling the needs of AI bias mitigation, making it imperative to interrupt and reform relevant labor practices before these practices scale further.

As T&S professionalizes, it is critical to address these continuing inequities, ensure clearer fair labor expectations, and support ongoing innovations in tooling that can mitigate psychological harm for the T&S community. Companies can put in place more stringent efforts to shield individual staff driving T&S internally and externally from public attack and can also support staff with additional security measures (physical and digital). They could also develop more stringent standards for contracting companies and create stronger systems that connect frontline content moderator expertise to HQ-based teams, given the deep analytical capacity and cultural context frontline moderators can offer regarding how harms are propagating within a certain community or language.

It is important to note that the **risks facing T&S practitioners extend to another key community of practitioners and moderators. Activists, researchers, and journalists often serve as first responders for their own constituencies.** These experts may be directly connected to individuals or communities who are facing extreme risk or harm, and may be powerless to help even when they have built trust or developed partnerships with individual companies. A common expectation that civil society and academic reporting be public (and made under an individual's byline) also increases risks to researchers, particularly for researchers or activists affiliated with marginalized communities already under attack. Activists, researchers, and journalists face equal or greater personal threat, harassment, and danger for their work on T&S issues, but enjoy variable access to formalized protections—and often no access at all. They may not even be able to seek protection from abuse on the very platform they are researching.

As one external expert stated in a task force convening, “T&S workers have the hardest job on the internet.” Working consistently at the heart of T&S dilemmas requires a level of resilience that most humans cannot sustain. It is imperative that this truth be recognized, acknowledged, and addressed continuously as online spaces shift, evolve, and expand

KEY FINDING 4

LEARNING FROM MATURE, ADJACENT FIELDS WILL ACCELERATE PROGRESS

Just as other sectors bring crucial expertise to the challenge of building healthier online spaces, so, too, do more mature fields. **One fundamental limitation of the current T&S field is how closely it hews to the culture, language, and incentives of US technology companies.** Such a corporate-centric framing impedes the creation of a more porous, generative relationship between companies and the wider range of stakeholders (including policymakers) who can offer critical insights and beneficial approaches for tackling complex harms and identifying unforeseen risks. This is particularly true of many civil society organizations, whose missions are often based on promoting and protecting “digital rights” rather than “trust and safety.” **As a basic example, “user safety” is a foundational concept and term for T&S practice. Outside of T&S, “user” is hardly a compelling way to describe a human being.**

Even as T&S practitioners strive to develop a more specific and standardized lexicon for the field, the lexicon itself will not translate with ease, either linguistically or normatively, into a vast range of cultures or contexts. (The same can be said of similar formations that have been driven by academia and civil society: “ethical” tech and “responsible” tech equally lack a normative footing across cultures and languages.) **Task force members highlighted the following fields as offering fundamental insights that should be incorporated more intentionally into debates and innovation around T&S as that field emerges.**

CYBERSECURITY

Cybersecurity¹¹ is a young field that has matured from being insularly technical to more multidisciplinary and multisectoral. It is often cited as a possible model for T&S’s evolution because both fields are composed of a diverse array of stakeholders focused on rapidly evolving technical and social disciplines while serving the needs of business and society.

¹¹ For a much more in-depth analysis of the intersection between evolutions within the Cybersecurity industry and T&S, please see [Annex 6: Learning from Cybersecurity, Preparing for Generative AI](#).

Understanding the main levers that supported the maturation of the cybersecurity field can offer insight into developments that could mature the T&S field more efficiently.

The cybersecurity community has made meaningful strides in the past decade in furthering education, inclusion, professional training, and research. This includes expanded educational [opportunities](#) and [certifications](#), focused efforts to build a [younger and more diverse talent pipeline into the community](#), and the creation of [governmental guidelines to develop the cybersecurity workforce](#). Creative, [team-based and immersive learning programs](#) have also taken root. Cybersecurity also promotes knowledge sharing through [journals](#), [conferences](#), and organizations such as [Information Sharing and Analysis Centers \(ISACs\)](#).

The vendor community has galvanized investment, publications, benchmarking and competitive progress (albeit sometimes unhelpfully through threat inflation, overtechnicalization of concepts, etc.). Cybersecurity-focused [journalists](#) have demystified the field for a broader audience by connecting the dots between cybersecurity and other key areas like national security and business, at least in the United States and Europe. Hackers have also helped structure the cybersecurity field.

In addition, many governments invested heavily in training people, developing policies, creating organizations, and passing legislation dedicated to cybersecurity. (Although this was facilitated by greater normative alignment between the cybersecurity field and government, and between governments, than the T&S space enjoys.) The development of sophisticated methodologies for characterizing vulnerabilities and malicious activity, best practices around various [methods of security disclosures](#), [bug bounty programs](#), and other non-remunerative disclosure mechanisms have all helped develop the cybersecurity field.

All of the examples above can serve as models for accelerating the development of a T&S field. For example, T&S could benefit by investing early in solving for weaknesses that cybersecurity has worked to overcome. Cybersecurity has [struggled](#) to make cybersecurity narratives accessible to nonexpert communities, and it is only in recent years that a long-standing “blame-the-user” narrative has begun to shift to a secure-by-design approach that emphasizes that primary responsibility for safeguarding users lies with platforms. While the advent of cyber insurance addressed some cyber risk, it has not driven companies to improve their cybersecurity as much as policymakers hoped. Moreover, while civil society, law enforcement, journalism, and researchers can and have served the same constructively adversarial function that hackers have within cybersecurity, they are not yet connected to the T&S practitioner community in the same fashion. Finally, T&S practitioners should not blindly follow in the footsteps of cybersecurity as a Global North-dominated field. Although Global Majority representatives play an active role in certain high-profile [commissions](#) and at the [United Nations](#), they do not drive the allocation of resources globally.

Moreover, in many countries, state-led cybersecurity action and agreements regarding cybersecurity have remained inaccessible and opaque to a broader community of stakeholders. National security and cybersecurity claims have frequently shielded contracts from scrutiny or oversight, for example, and have also been used as a pretext to bar civil society, researchers, or journalists from accessing information regarding potentially rights-violating activities conducted in the name of cybersecurity. The T&S community can learn from this example by building and protecting transparent (or at least not entirely opaque), multistakeholder processes from the outset as a de facto standard for the field.

HUMAN RIGHTS

International human rights law is a field benefiting from seventy-five years of evolving debate, language, norms, frameworks, and implementation models, including the UN Guiding Principles on Business and Human Rights, many of which have been contextualized to the digital environment at [global](#), [regional](#), and domestic levels. **General consensus within the task force supported the finding that greater interoperability**

between T&S and human rights could serve to strengthen both fields, identify new pathways for achieving T&S goals, and improve T&S's ability to narrate its aims more clearly with a wider community of stakeholders.

As companies face a new era of regulatory requirements and compliance frameworks, economic and legal pressures may incentivize companies to make the regulatory floor their T&S ceiling, and to shift investments away from more proactive or innovative approaches to building T&S (such as prosocial product-design methodologies or expanded multistakeholder engagement). Human rights impact assessments (HRIA) and due diligence assessments can help protect space for key T&S equities and maintain a forward-looking and expansive focus that audits are not necessarily structured to provide. For example, when a video-streaming platform published the results of its [first independent HRIA](#) in April 2023, [multiple findings dovetailed exactly](#) with key T&S concerns. The platform noted as a key takeaway that, “despite [our] lower risk profile today, as we build and grow, we must continue to acknowledge that not every user has the same experiences, and that some groups are particularly vulnerable to human rights risks and abuses on our service. This is important as we consider whether to expand globally into new markets, and how core product decisions may affect [our] evolution as a service.”

At a time of intense debate and policymaking focus around key T&S issues such as children’s safety, using a rights-centric framework can help establish a foundation for normative debate and key trade-offs by positioning safety priorities within a broader backdrop of long-standing other rights (such as privacy, or access to information).¹² [Rights-based self-governance mechanisms](#) have also played a meaningful role in driving multistakeholder consensus that can then inform coherent policies and regulations in the future.

International human rights law has its own limitations as a field and a framing for T&S. Implementation approaches vary depending on the jurisdiction; voluntary principles lack strong enforcement mechanisms; participatory inclusion of the parties impacted by a policy or product are not guaranteed; and state-centric models can offer drawbacks at a time of increasing autocracy, among other issues. In addition, it is critical to note that within the United States, civil rights are a far more powerful foundation than human rights for protecting and promoting the rights of marginalized and disenfranchised communities, especially vis-à-vis US companies. In many other countries, fundamental human rights are the foundation of domestic law and must also be read into any domestic law or regulation related to the digital environment.

Numerous members of the task force cited the high-level normative basis of human rights analyses as a weakness that must be balanced with a parallel mapping of the concrete risks being created by a particular service. **Both the clear identification of harms or risks and the clear identification of implicated rights are necessary inputs to solutions-oriented discussions internally and externally.**

ADDITIONAL KEY FIELDS

Exciting and important corollaries exist between broad T&S goals and needs and a range of other fields. Lessons could be pulled from:

- ▶ Finance, particularly with regard to the evolution of global standards, statutes, and tooling to combat money laundering; the development of a strong media presence in the industry that promotes accountability as well as education across stakeholders; and how national financial intelligence units have attempted to lower the reporting bar for risk information.
- ▶ Public health, with a particular focus on [how public health could serve as a model for new types of technology governance](#), as well as how the field has navigated knowledge shar-

¹² For a more in depth analysis of children’s rights, see [Annex 3: Respecting Children as Rights Holders](#).

ing and the safe aggregation of sensitive data for large-scale, longitudinal, and cross-border research, innovation, and accountability mechanisms.

- ▶ Urban planning, citing among other key examinations the work of [New Public](#).
- ▶ Civic technology and “[GovTech](#),” with a particular focus on how civic technologists and government technologists have invested in building new and interesting forms of [deliberative polling](#) and [democratic governance](#), as well as best practices for publicly funded and community-driven online spaces; and how these lessons could inspire [new approaches](#) to long-standing questions of T&S governance and policymaking within services.
- ▶ Advertising and “Ad Tech,” with a particular focus on how advertisers have leveraged their collective market power to standardize requirements for brand safety through the creation of initiatives like the [Global Alliance for Responsible Media](#) and the [Oasis Consortium](#), as well the development of measurement practices facilitating verification of and optimization away from harmful content.

This list is hardly dispositive. Rather, it serves as a reminder of the breadth of work being done across the broader digital ecosystem that could, coupled with the increasing emergence of T&S, serve to power a brighter and more trustworthy digital future.

KEY FINDING 5

THE GAMING INDUSTRY OFFERS UNIQUE POTENTIAL FOR INSIGHTS AND INNOVATION

→ Ongoing global debates over online spaces tend to focus on major social media companies like Meta, Google, or Twitter. This inevitably shapes discussion and ideation around approaches to content moderation, trust and safety, and even future technology. Gaming has long served as a significant piece of the growing digital environment; it is estimated that three billion people around the world play digital games, with a projected market value of more than \$300 billion by 2026. Historically, though, gaming has been isolated from policy communities focused on internet governance, social media, and “big tech” issues, and that has resulted in a lack of appreciation for the gaming industry’s long-standing market share, geopolitical impact, technological innovation, and connection to the rest of the information ecosystem. **Understanding this industry¹³ is an increasingly important element of understanding where and how digital spaces might evolve, and that means examining not only games themselves, but also the industry’s ownership, incentives, and business models.**

Much of the emerging immersive technology is being developed through the gaming industry, and active experimentation is taking place with applications of distributed technologies and AI. As immersive technologies become more pervasive they are likely to be grounded in the gaming ecosystem, and may also carry with them many of the challenging dynamics games have long grappled with, including hate speech, bullying, illicit activity, and harassment. **There are lessons to be learned from the industry’s successful and less successful approaches to content moderation, trust and safety, and product design.** Games have also long existed as multimedia interactive spaces that commingle real-time mixtures of audio, video, and text components as a key feature: one that will define online spaces more and more in the future. With the increasing popularity of VR games and applications, companies are focusing on developing new safety features to protect users in

¹³ For a deep dive into the gaming ecosystem, see [Annex 4: Deconstructing the Gaming Ecosystem](#).

these immersive environments and also bring long-standing expertise to bear regarding the pros and cons of achieving different levels of fidelity within a particular digital environment. Efforts to improve the real-time monitoring capability in privacy-respecting and less data-intensive ways will have applications for numerous industries. Finally, gaming is already grappling with the increased aperture of user-generated content as a threat model in the age of GAI, as barriers to content creation drop dramatically and monetization models rapidly open up to a broader array of individuals and incentives.

Another unique element of gaming is the industry's expertise in designing for the intentional inclusion of children, including those younger than thirteen, as well as adults. In addition, in recent years some firms in the gaming industry have added a normative frame to game development, pioneering prosocial approaches: more intentional and proactive design methods that preemptively shape and encourage healthy and inclusive play patterns at all ages. These methods pull from best practices in design, psychology, sociology, and more, as well as case studies from earlier multiplayer games. The gaming world has also leaned into the idea of enabling unique rules and norms for unique spaces, set and enforced by communities. Better understanding the mechanisms, benefits, and drawbacks of all these approaches would serve a broad community outside of gaming.

Finally, **the gaming ecosystem is global in scope and mirrors many of the broader debates over current questions of critical technology, investment, ownership, and norms.** Many of the world's largest gaming companies are headquartered in the United States and Europe, with major companies also found in Canada, Japan, and South Korea. Many of these dominant companies have received substantial investment from Chinese and Saudi Arabian government-backed firms. Indeed, both countries are placing significant emphasis on building ownership stakes in foreign gaming companies and increasing the reach of their own industries in the lucrative market. In addition, the games industry is trending toward consolidation as major industry players buy up indie and midsize game-development studios, and as tech giants such as Microsoft seek to acquire gaming giants like Activision. This trend mirrors and overlaps with the evolution of existing, major social media and tech platforms (e.g., Meta acquiring Instagram, Google acquiring YouTube). As gaming technologies become core components of the future web, understanding the impact such investments may have on market incentives, content, product, and trust and safety practices will be important. As more of the gaming ecosystem and social media-dominant digital spaces converge, questions of which regulations and oversight bodies might apply will also emerge as an important area for clarification.

KEY FINDING 6

EXISTING HARMS WILL EVOLVE AND NEW HARMS WILL ARISE AS TECHNOLOGIES ADVANCE

Where known risks exist in traditional online spaces, it is inevitable that the same risks will migrate to any online spaces powered by emergent (or newly popular) technologies. From the recent, more widespread adoption of federated spaces (see below), to the emergence of eXtended reality (XR) platforms and increasingly metaversal forms of gathering, to the rise of generative AI—even as known risks and harms travel—the policy, product, and tooling solutions that have been developed for more traditional online spaces may not be applicable or even technologically feasible. In addition, entirely new sets of risks may emerge with new technologies that are not yet adequately understood, as will new opportunities.

FEDERATED SPACES

The emergence and growth in popularity of federated¹⁴ social media services, like Mastodon and Bluesky, introduces new opportunities, but also significant new risks and complications. While federated services continue to be dwarfed in size in comparison to platforms like Facebook and Twitter, the steady rise in their adoption warrants further attention and study. These emergent distributed and federated social media platforms (aka the “fediverse”) offer the promise of alternative governance structures that empower consumers and can help rebuild online spaces on a foundation of trust. Their decentralized nature enables individuals to act as hosts or moderators of their own “instances,” increasing user agency and ownership. Platform interoperability ensures users can engage freely with a wide array of product alternatives without having to sacrifice their content or networks.

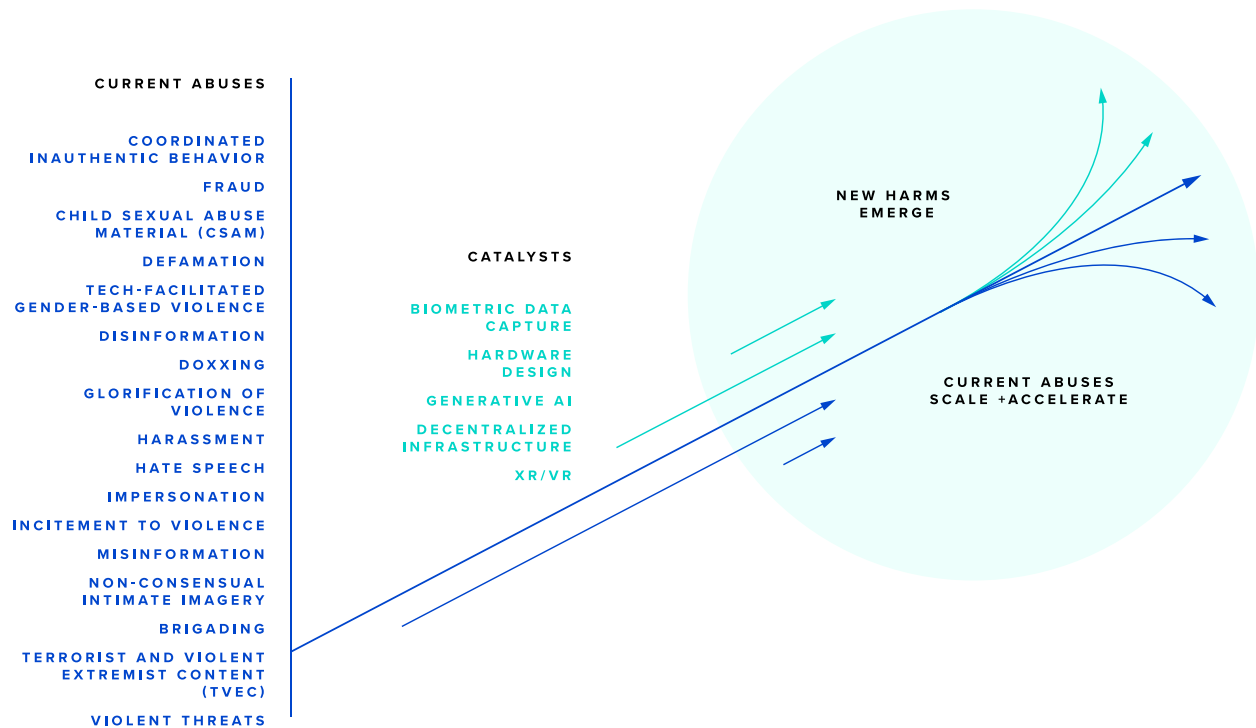
¹⁴ For a deeper dive into this topic, see *Annex 5: Collective Security in a Federated World*. Broadly speaking, the “fediverse” is a catch-all term for a wide array of distinct products, services, and platforms that interconnect using a set of shared communication protocols such as the W3C standard [ActivityPub](#) or the still-in-development [Bluesky AT Protocol](#).

Federated spaces have many of the same propensities for harmful misuse by malign actors as mainstream platforms like Facebook and Twitter, while possessing few, if any, of the hard-won detection and moderation capabilities necessary to stop them. Each instance of a federated service can choose for itself what its governance approach will be. Community standards, content moderation, user reporting, and protecting against large-scale or coordinated campaigns of harassment or disinformation—even within an individual instance—require a broad array of technical, institutional, financial, and logistical competencies that federated spaces are not currently designed to support.

Across instances, it’s challenging for instance moderators to engage with each other in a structured way to counteract shared threats. While decentralized community governance has had notable successes on platforms like Wikipedia, lack of shared norms and standards across instances impedes the adaptation of those collaborative practices to the fediverse. Indeed, absent the financial support that goes along with centralized, corporate social media, few parts of the fediverse have been able to successfully marshal the human and technological resources required to successfully execute proactive, accurate T&S services at scale. **The unit economics of toxic or manipulative behavior are currently skewed firmly in favor of bad actors, not defenders.** They also incentivize the creation of closed communities with a high degree of cultural alignment, which not only offer extraordinary opportunity for resilience and community building, but also foster communities that spawn radicalization, hate, and other toxic byproducts. Adding to this challenge is the existing uncertainty regarding emerging regulation and how it will be applied to federated instances.

Many of the above challenges are (at least partly) solvable product, logistical, and engineering challenges. Others are deeply ingrained cultural behaviors that will take considerable time to change. All will require sustained focus, attention, and innovation to address.

COMMON TYPES OF ONLINE ABUSE



IMMERSIVE SPACES

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 concerns also loom large. For example, the volumes of data collected and traffic sent as part of gaming platforms are of interest to companies, governments, and potentially criminal actors as well. XR environments may be centralized or decentralized as well, and the risks and opportunities present in those respective environments (as narrated above) reflect those shared by non-XR spaces.

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 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, and 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 a user can be ascertained by their behavior within the ecosystem, and that can scale up or down across a range of hardware or platforms depending upon any use. 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. When creating policies and terms of services to moderate users, services will have to consider the unique ways users interact with a technology that blurs the divide between virtual and physical worlds, along with the unique affordances of technology. This means adapting policy to focus on behavioral interactions in addition to speech-centric interaction, and developing tooling to support that shift.

CONTENT & CONDUCT MODERATION

The content moderation issues debated in the T&S space today apply to XR as well, but **tooling norms and regulations (which are already quite complicated, fragmented, controversial, and quickly evolving), will need to evolve to properly address emerging technological contexts**. Moderation of social VR and audio/chat functions is particularly difficult and can be costly. Recently, moderation companies have been investing 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). As GAI inevitably lowers the barrier to creating synthetic media, it is foreseeable that deepfakes and additional forms of audio- and video-based impersonation—which were already a growing problem before GAI—will increasingly pervade XR spaces, creating new opportunities not only for harassment and disinformation but also for financial fraud.

USER STANDARDS AND SAFETY

Though video game and social media addiction have been more widely studied than VR applications, consumer safety concerns have emerged for the latter in the past couple of years: from eye strain to the psychological impacts of being physically or sexually assaulted in a virtual world. Specific risks to child safety will need to be considered and negotiated as adoption increases; indeed, Meta recently opened Horizon World to teen users in the United States and Canada and placed specific limitations on their accounts. **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” also run the risk of being even more harmful in immersive environments, although innovative mitigations are already being piloted. In addition, the normalization of chance-based monetization systems (sometimes called “gambification”) in games is raising important questions about safety from commercial exploitation and from technologies specifically designed to foster compulsive behavior or even addiction among players.

PRIVACY

It is **still not clear how privacy will be conceptualized and ensured in XR environments** given their interoperability requirements and the sheer amount and range of sensitive data required to support VR and even augmented reality (AR) environments. Particularly as XR hardware continues to evolve and become more standardized, user security and understanding of risks, opportunities, and assumptions of use will be important touch points for companies, regulators, and watchdogs alike. In addition, as companies and researchers experiment more with using on-device computational capabilities, current data storage and processing standards and risk models are likely to evolve dramatically.

EQUITY AND ACCESS TO XR TECHNOLOGY

If developed and distributed correctly, **XR has enormous potential to help increase accessibility**. XR technology enables more equal access to virtual experiences and content, promotes inclusivity, and improves the user experience. In order to aid the positive benefits, stakeholders **need to keep engaging in discussions about diversity, equity, and inclusion; international development; and education**. This should happen alongside broader conversations about access to underlying technologies (e.g., 5G) necessary for inclusive and safe adoption in communities traditionally excluded from early access.

GENERATIVE AI

GENERATIVE TECHNOLOGIES AND THE INDUSTRY OUTLOOK FOR TRUST AND SAFETY

Generative AI¹⁵ refers to powerful algorithms that can produce or generate text, images, music, speech, code, or video. These algorithms rely on large language models, consisting of vast artificial neural networks, and are trained by consuming and processing large amounts of data. While not a new technology, the wildly popular release of ChatGPT and DALL-E at the end of 2022 catapulted GAI and LLMs into the public sphere. Leading technology companies ranging from Google to Meta to newer AI-focused entrants, such as OpenAI and Anthropic, have invested heavily in developing their own LLMs and associated products for public use. Governments, investors, and innovators alike have refocused their attention on these models and the products they power given GAI's potential to reshape society.

GENERATIVE AI: FRIEND OR FOE TO CONTENT MODERATION?

Generative AI changes the nature of influence operations online and the moderation of illicit content by reducing the financial cost, time, and technical expertise required to produce mass amounts of hyper-realistic harmful content and potentially spread it at scale. Automating the production of fraudulent content, misinformation, spam, influence operations, and other forms of illicit online behaviors through GAI results in content that is more convincing than previous forms. Increased volume of deepfakes not only risks flooding trust and safety systems with exponentially greater quantities of content that will need to be monitored, but also injects greater quantities of hard(er)-to-detect forms of high quality (and potentially harmful) fake content into the system.

¹⁵ For deeper and/or additional analyses of GAI, please see [Annex 1: Current State of Trust and Safety](#); [Annex 2: Building Open Trust and Safety Tools](#); [Annex 4: Deconstructing The Gaming Ecosystem](#); and [Annex 6: Learning from Cybersecurity, Preparing for Generative AI](#).

LLMs may also change the nature of influence campaigns. Previously, disinformation campaigns focused on easier-to-generate artifacts such as text and image. It is not yet clear what a targeted disinformation campaign might look like in the era of easily developed video and voice. Put simply: people do not yet have the reflex for critical consumption for video and images as they have for online text-based content.

Toxicity and abuse online are not simply matters of content-based harms, but can also involve highly nuanced actor and behavior-based challenges, which current LLMs may be less equipped to solve. Furthermore, LLMs are sycophantic and have no internal model for truthfulness of factuality. Systems deployed today also do not learn in real time, instead being trained on data up to a cutoff point due to the time-consuming nature of training. Models must be regularly trained and realigned as company content policies change. There are also additional privacy risks in AI-powered harvesting of content, especially as companies collect and store more user data and expand red-teaming exercises to include an ever-widening array of individuals (thereby, increasing the risk of leaks and abuse of data and LLMs, etc.).

Additionally, all currently existing LLMs are built from content ingested from the open web. This means that not only racial, cultural, and religious biases, but also illegal behaviors, toxic content, hate speech, and even personally identifiable information (PII) are all present and accessible within the knowledge-base of these systems. Because a biased LLM is unable to detect and remove its own bias, some AI providers are experimenting with using pre-cleaned, PII-scrubbed, EU regulation-compliant,¹⁶ and detoxified datasets to retrain and fine-tune LLMs in order to remove these unwanted toxicities from the LLM itself. Until the content within the LLMs themselves has been moderated, they are prone to the age-old technology aphorism: “garbage in, garbage out.”

While GAI drastically changes the scale and speed at which malicious online behaviors occur, GAI also might serve as a tool for trust and safety professionals looking to mitigate these very same harms through data curation, model training, and postdeployment evaluation of existing content. Examples of this include automatically attaching warning labels to potential generated content and fake accounts; improving vetting, scoring, and ranking systems; creating high-quality classifiers in nonmajority languages; and quickly moderating spam and fraud through GAI.

¹⁶ Specifically, this refers to datasets compliant with the EU's General Data Protection Regulation

KEY FINDING 7

SYSTEMIC HARM IS DRIVEN BY MARKET FAILURES THAT MUST BE ADDRESSED

One fundamental point of consensus across the task force was that **risk and harm are currently set to scale and accelerate at an exponential pace, and existing institutions, systems, and market drivers cannot keep pace.** Industry will likely continue to drive these rapid changes, but will also prove unable or unwilling to solve the core problems at hand. Major regulation from the European Union and elsewhere is creating new incentives and driving new practices across the technology industry that are shifting markets and existing practices, but governmental action has perennially proven incapable of keeping pace with emerging technology (unless that action has been to censor, surveil, block, or otherwise violate fundamental rights and freedoms). The task force's focus on conducting systems-level analyses highlighted three areas in particular that merit deeper examination based on how they impact the incentives structures that truly govern the digital space. **Until investments in reactive and proactive T&S are established as a requirement for doing business or a de facto generator of long-term value, the incentives structures necessary to ensure better, safer online spaces will continue to fail users—and societies.**

MEASURING T&S IS A MEANINGFUL CHALLENGE

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 within companies. 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 tools¹⁷ that can be used

¹⁷ For a deeper dive, please see [Annex 1: Current State of Trust and Safety](#).

across different companies to define whether a company’s investments in trust and safety are a driver of long-term growth, either by adding value to the product, improving customer experience, burnishing the platform’s credibility, protecting revenue generation, or otherwise. Some notable progress is being made in this regard.

The absence of maturity models also continuously undermines T&S forecasting, investments, and prioritization. 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 the safety of users. In addition, the investments a company needs to make in T&S to protect the company’s own reputational risk (another common means of evaluating T&S costs) 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 midsize online service providers found that for midsize 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.”

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 or CEO to know if the company’s T&S team is excelling or lagging against a standard industry expectation. T&S is not amenable to conventional performance metrics such as objectives and key results (OKRs), and requires a range of new metrics that can 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 the tech sector.

Perhaps most importantly, **few external incentives currently force a C-suite or board to care about T&S.** Even where T&S team performance can be measured, that does not guarantee measurement of harm across a platform. Even if harm can be measured across a platform, senior executives can ignore those findings at their discretion. The emergence of new and widespread **regulatory requirements will fundamentally reshape how companies evaluate investments and forecast costs, and can help create some external pressures—but more is needed.**

EMERGING REGULATION IS ALREADY A MARKET DRIVER FOR T&S

As is often the case, government regulation in the tech sector has followed, rather than led, the bulk of industry action on T&S. This means much of the current regulatory conversation is responding to the teams, skills, tools, and capacities companies had already created in response to incentives other than digitally focused laws.¹⁸

However, as public concern has mounted over individual and societal-level harms that are scaling at a break-neck pace, and as core societal functions have grown dependent on privately owned platforms, governments have increasingly begun to step in. In some instances, regulations are aimed at increasing corporate accountability and protecting citizens’ rights; in other instances, regulations have been designed to increase surveillance and increase political control within a country’s borders. Across the board, this proliferation of competing and sometimes contradictory rules is making it difficult for companies to navigate the numerous markets in which they now operate.

Many countries have approached regulation in a piecemeal manner—passing laws focused on specific content concerns, child safety issues, competition, or even product features like algorithms. The EU’s Digital

¹⁸ Existing laws in key areas such as privacy, expression, child safety, terrorism, fraud, criminal activity, and intellectual property (among other areas), have long played a meaningful role in driving T&S decisions.

Services Act, discussed above, and Digital Markets Acts (DMA) are notable for harmonizing laws across twenty-seven member countries, which—given the respective power of that economic market—will also be a primary driver in consolidating industry compliance priorities and funneling heretofore voluntary approaches into a more standardized legal enabling environment. This matters because **the DSA and DMA will establish the foundation for what data and information companies are required to share, with whom, and about what, as well as how companies contemplate and manage systemic risk.** Elements of each of these requirements are actively being considered by a number of other governments, which are likely to match at least some of the standards developed through these European laws.

Numerous stakeholders have given significant, serious attention to the content and context of emerging regulations, as well as to [tracking regulations as they emerge](#). The impact that regulation will play in reshaping the T&S field and its incentives, however, is less widely understood outside of industry, and merits attention from a wider range of stakeholders. As a tightening economy pushes the tech sector and private finance to make cuts and minimize investments, companies are reshaping T&S investments, both internally and through vendors in order to support compliance. This has included widespread layoffs of T&S teams, as well as a reputed move toward bringing in new hires from industries with a stronger basis in auditing and compliance processes, such as finance. Internal investments in many traditional T&S areas, among them risk assessment, due diligence, documentation of enforcement mechanisms and metrics, and responses to governmental requests for sensitive data, are shifting as industry pivots to respond.¹⁹

T&S vendors are also adjusting their offerings to support transparency reporting and other compliance workflows, and there are indications that a new start-up market is emerging to support “T&S as a service.” On the plus side, a thriving vendor market could allow companies to take on a wider range of T&S functions due to increased access to external expertise (technological, contextual, linguistic, or otherwise) and improve the maturity of the field. By the same token, the rapid expansion of a vendor market lacking standards for vetting or due diligence may primarily serve to help companies externalize their T&S risks without taking on significant responsibility for deepening their T&S expertise in-house or understanding where their service might be creating risk or generating harm. Finally, shifts to even more advanced AI-based content moderation tools may create the impression that human beings will no longer be needed to support this function. The true answer is that human moderation will remain a critical component of T&S, but new tooling may shift where human moderation is focused and prioritized.

Members of the task force specifically warned that **while the EU standard could increase industry focus on trust and safety policies, practices, products, and tools, it could also divert attention and resources from the most vulnerable communities and markets—particularly non-English language ones. Another widespread concern: if compliance replaces problem-solving, it establishes a ceiling for harm reduction, rather than a floor founded in user and societal protection.** Compliance regimes can calcify reactive practices, diminish C-suite appetite for innovation and proactive approaches to improving T&S, and undermine teams that are seeking to solve the underlying problems enabling harm. Another risk identified was a move away from assessment frameworks, which are by nature forward-looking, and toward audit frameworks, which are focused on current and past practice and narrowly delimit a scope of review. At a moment when information sharing is critical to the expansion and professionalization of T&S, many experts worry that they will face even greater barriers to tracking or sharing any information beyond that which is mandated.

Finally, while there is no question that self-governance alone has been insufficient to ensure adequate attention to T&S, **technology will always move faster than regulation. Many task force members cited the value of voluntary/self-governance initiatives in supporting knowledge exchange and the shaping of norms**

¹⁹ The United States is notable for the impact and incoherence created by subnational laws, as well.

and best practices within emerging technologies, and expressed their hope that investments would not move away from supporting those collaborative mechanisms. In addition, task force members highlighted that regulation can—where carefully constructed—play a powerful role in preventing races to the bottom. Among other measures, transparency requirements and mandates for researcher access can empower researchers, civil society organizations, and governments to better understand the policies and practices necessary to build healthier online spaces at the speed the internet requires.

THE ROLE OF VENTURE CAPITAL HAS BEEN UNDEREXAMINED

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. Privately funded companies face little pressure from investors to demonstrate or design a T&S strategy, and T&S vendors have, with some exceptions, historically struggled to attract significant and continued investment compared to other technologies. Instead, many have been acquired by larger companies seeking to bring capacity in-house. One result has been that VCs remain unclear on the market segmentation and exit potential for T&S vendors.

In addition, investors and executives have failed to connect durable value generation with investment in T&S practices. This connection has recently been illuminated by the reported decimation of Twitter’s revenue stream and its increased risk of significant fines most likely due (in part) to moves that weaken T&S practices, such as withdrawing from the EU’s voluntary Code of Practice on Disinformation. **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.

During this historically low-period of VC fundraising, **building a cohesive, systems-level T&S strategy may include changing the incentives of VCs.** This could include campaigns to raise awareness with their important limited partners (LPs) as well as direct work with VCs to illuminate the virtues of providing meaningful T&S portfolio services to investees, especially as regulatory requirements increase and GAI investments skyrocket. Given the dialogue from the public and private sector around how to build AI companies responsibly, there is an opportunity to ensure that T&S considerations are included in the frameworks and resources that are developed for technologists. It may be equally important to explore existing limitations to VC-funding models in order to triangulate where other forms of investment or resourcing will be more effective or sustainable.

KEY FINDING 8

PHILANTHROPIES AND GOVERNMENTS CAN SHAPE INCENTIVES AND FILL GAPS

→ **Philanthropies can play a transformational role in helping to fill systemic gaps the task force identified.** From catalyzing research into market drivers and sound business practices to funding research to driving collaboration, philanthropy is ideally suited to inject resources into the broader ecosystem and expedite forward movement. In areas where industry is most likely to pull back investments over the coming years, or where extreme inequities must be balanced in order to support safer, better online spaces in the future, philanthropy is ideally positioned to respond. This can include seed funding to support the scoping and negotiation necessary to set the stage for large-scale endeavors, such as independent governing bodies that might eventually be supported by industry, international governmental bodies, or the public sector.

Beyond regulation, governments can play a constructive and creative role in supporting independent research, deploying foreign assistance funds, seeding innovation, documenting and making public their own expert practices in building safe public spaces online, establishing or supporting educational programming, and designing proactive policies and funds to support industry (particularly small and medium enterprises) to take on best practices that might not otherwise be rewarded by market dynamics. Governmental actors charged with engaging the tech sector could work actively with counterparts in public interest technology, digital public infrastructure, digital public goods, and digital services to understand where public investments in tooling, product design, and policy development could also be given further reach.

KEY RECOMMENDATIONS

Acknowledging the unique capabilities of the philanthropic sector, the task force focused particularly on identifying opportunities for philanthropic investment that could fill systemic gaps and catalyze novel and creative pathways to achieving systems-level change.

The task force urges significant and immediate investments designed to:

- 1 Craft and implement initiatives to target market failures and incentives gaps.**
- 2 Accelerate the maturation and professionalization of trust and safety as an independent field.**
- 3 Break down knowledge silos and share information and expertise.**
- 4 Protect and grow the enabling environment necessary to innovate more trustworthy useful online spaces.**
- 5 Expand investment in proactive, future-facing research and initiatives.**

CRAFT AND IMPLEMENT INITIATIVES TO TARGET MARKET FAILURES AND INCENTIVES GAPS

- 1.1** Fund market research that connects trust and safety (T&S) more naturally to market drivers and helps fill known market gaps. Examples include: model metrics to measure return on investment for T&S teams and tools; case studies on the business impacts of T&S; studies defining and scoping the current and potential size of the T&S market; more extensive public research on norms for T&S expenditures within platforms operating at different scales and revenue levels, among other publications and projects.
- 1.2** Support efforts to connect the advertising industry's ongoing research and analysis with the broader T&S ecosystem, including research and analysis on the impact of affiliating with high-quality, brand-safe online content.
- 1.3** Support studies and public campaigns documenting and calculating the cost of noteworthy T&S failures or underinvestment.
- 1.4** Raise awareness of T&S tools, implications, and approaches with limited partners in the venture-capital community, and increase pressure on firms to provide T&S training and support as a portfolio service.
- 1.5** Explore alternative and mixed funding models for infrastructure gaps the market struggles to fill, and fund studies to clearly outline tools and needs the market cannot bear.
- 1.6** Focus existing and developing regulatory frameworks on incentive gaps related to revenue-generation models, systemic harms, and knowledge imbalances.

ACCELERATE THE MATURATION AND PROFESSIONALIZATION OF TRUST AND SAFETY AS AN INDEPENDENT FIELD

Promote and Expand Collaboration and Knowledge Exchange among Trust and Safety Practitioners

- 2.1** Provide sustained support to publications, conferences, communities, and convenings—especially Global Majority-run T&S events outside of the United States and Europe—that allow T&S practitioners to engage outside of their own companies and teams, and exchange best practices with a broader professional community in trusted spaces.
- 2.2** Build pathways for collaboration and field building between T&S practitioners in adjacent industries (like gaming or gaming-related social media) and from teams that do not call themselves “trust and safety” (e.g., human-rights teams in some companies) with the growing and formalizing T&S field.
- 2.3** Support the creation of more T&S teams in key industries, such as gaming.

Protect the Wellness and Resilience of T&S Practitioners, Particularly Content Moderators

- 2.4** Call for the use and development of new moderation tools that enable interventions, such as image blurring, to mitigate harm to human moderators.
- 2.5** Implement workplace-wellness programs to address the needs of those exposed to harmful content.
- 2.6** Provide digital protection services to key T&S employees to keep personal information, such as home addresses, off easily accessible sites, and monitor harassment generated by enforcement decisions.
- 2.7** Better integrate frontline content-moderator teams and expertise with headquarters-based staff at companies, and establish stringent industry standards for companies contracted to provide content moderation, artificial-intelligence (AI) model training and testing, and other related support to ensure they are appropriately compensated and protected from harm.

Invest in Building a Diverse T&S Expertise Pipeline

- 2.8** Fund the creation of model T&S curricula and other educational programs for high-school, community-college, university, and graduate-level students across computer engineering, political science, history, user-experience design, product development, and other related disciplines.
- 2.9** Establish university courses, ensuring such courses are funded and supported in countries outside of the United States and Europe, and in a diverse range of educational institutions.
- 2.10** Create professional certifications for various T&S-focused skills (e.g., data science, content moderation) and knowledge areas (e.g., bullying and harassment, child sexual abuse material), modelled on approaches from other industries such as the [SANS Institute](#) in cybersecurity, and ensuring such certifications are supported in countries outside of the United States and Europe, and through a diverse range of educational institutions.
- 2.11** Support the development of inclusive hiring pipelines from under-represented or nontraditional professional backgrounds into T&S roles, including through job fairs and recruiting events targeting specific groups and professional communities.

Support Metrics Standardization Across the T&S Field

- 2.12** Develop a framework like that in [the cybersecurity world](#) to articulate the full range of roles, skills, and competencies across all sectors of society (including regulators, civil society, etc.) that comprise the T&S workforce.
- 2.13** Establish and promote voluntary standards, certifications, and transparency measures that T&S vendors can adopt to drive consistency and comparability in the vendor ecosystem.
- 2.14** Launch longitudinal studies to track whether professionalization strategies (such as certifications and higher-level coursework) endanger geographic and socioeconomic diversity among T&S practitioners over time.
- 2.15** Develop a common-harms rubric for use across platforms. This could draw inspiration from the cybersecurity world's [Common Vulnerability Scoring System](#) framework.
- 2.16** Publish an accessible guide for startups on a scalable approach to trust and safety at key junctures in company growth—whether moving through increases in user engagement, expanding to new markets, or other touchpoints known to create new T&S needs. Include available tooling at each stage, insights on build or buy decisions, and other known best practices.

BREAK DOWN KNOWLEDGE SILOS AND SHARE INFORMATION AND EXPERTISE

Apply Lessons from Other Industries to Common Challenges

- 3.1** Establish pathways for constructive adversarial T&S work, building upon work developed over years within ethical hacker communities. This could include adoption of T&S security disclosures, [bug bounties](#), and other mechanisms to incentivize the discovery and disclosure of systemic risks and vulnerabilities in policies and enforcement. Such programs should be geared toward creating an avenue for collaboration and discussions between companies, and within the broader community working to keep the Internet safe and open.
- 3.2** Develop one or more Information Sharing and Analysis Center ([ISAC](#))-like cross-platform information/threat-intelligence sharing organizations to facilitate information flow between and among companies on priority online harms. For example, an elections-focused model could be explored as a pilot.
- 3.3** Fund experiments exploring applicability of practices from related fields to one another. This might include applying prosocial design methodologies developed in the gaming world to traditional social media; moderation practices from community platforms like Wikipedia to others; or enforcement practices from interactive contexts to more traditional platforms.
- 3.4** Apply insights from the gaming world's experience with non-text-based and mixed-media contexts to other social and interactive digital forums. In particular, convene workshops, research, and experiments on how the known harms, tradeoffs and challenges in gaming spaces are likely to manifest in more immersive and mixed-media version of social, political, and other interactive platforms.
- 3.5** In addressing concerns over risks to youth online, apply a rights-based framework to their engagement, to ensure consideration of the range of relevant safety, privacy, and other protections.

Develop Stronger Participatory Inclusion Models

- 3.6** Support Global-Majority organizations in contributing and scaling tools, methodologies, indexes, frameworks, and other contributions to the theory and foundations of T&S regulation, policy, product design, and practices. This expertise is valuable for its contribution to addressing challenges faced by everyone online (as Global Majority countries are often the first to experience them), not just for context-specific case studies.
- 3.7** Support organizations that work with companies to develop products and policies that ethically and effectively account for and include marginalized, vulnerable, or particularly at-risk communities. This might include directly working with youth in the development of products targeting them; non-English or dominant-language-speaking communities in the user experience of a product expanding to new markets; or at-risk activists in the design or settings of policies affecting their ability to use digital tools.
- 3.8** Support the development of easy-to-access, engaging tools for improving awareness of T&S, such as games and easy explainers/primers.

PROTECT AND GROW THE ENABLING ENVIRONMENT NECESSARY TO INNOVATE MORE TRUSTWORTHY, USEFUL ONLINE SPACES

- 4.1** Provide flexible, general support to civil-society organizations and leaders working within, and on behalf of, their own communities to understand how technology is used, abused, and broadly impacts society. Ensure this network of organizations, particularly in marginalized communities in the Global North and those outside of the United States and Europe, is able to sustain and grow efforts to monitor, document, and inform companies, regulatory structures and processes, standards-setting bodies, governance forums, and other civil-society colleagues working to understand and improve the digital world.
- 4.2** Ensure funding to organizations, independent researchers, and experts to work, engage, and possibly begin creating tech policy and product hubs in emerging centers of regulatory and technological power around the globe.
- 4.3** Provide access to protection support for civil society, researchers, and policymakers who contribute to the success of T&S practices and the health of the Internet, and are often physically and digitally targeted for that work.
- 4.4** Ensure that regulatory provisions requiring consultation with external civil-society experts (on topics such as risk assessments, systemic harms, etc.) also account for the material support civil society needs to fulfill that role and provide such services.
- 4.5** Co-design usability improvements to processes and tools that end users utilize to engage with T&S teams, with particular attention on marginalized populations. This may include processes related to reporting, deplatforming, harassment, malicious flagging, and other matters.

- 4.6 Facilitate access to centers of power by global-majority organizations. This may include providing dedicated spaces for co-living/co-working in places like Brussels, San Francisco, or Washington, DC, as well as supporting legal and technical assistance with relocation costs, immigration and employment complexities, and sustainability models.
- 4.7 Fund research and pilot new funding models to support civil-society organizations working in contexts where foreign funding of rights/risk-based activities are increasingly monitored and curtailed.

EXPAND INVESTMENT IN PROACTIVE, FUTURE-FACING RESEARCH AND INITIATIVES

Encourage a Race to the Top

- 5.1 Accelerate small and mid-sized platforms' deployment of high-quality T&S policies, tools, and operational practices. This can be most readily achieved by establishing an independent, nonprofit entity—a T&S Tooling Accelerator—to develop, maintain, and grow new open-source tools, policies, and best practices; obtain existing tools (e.g., donated/licensed) from platforms and vendors, and package and distribute them free of charge or at a greatly reduced cost to participating platforms.²⁰
- 5.2 Consider establishing T&S-focused awards (or alternative recognition models) within key sectors—such as government, industry, media, academia, or civil society—that could identify promising new innovations within the T&S ecosystem, as well as consistent use of best practices. Such awards could be particularly important within government and industry.
- 5.3 As multiple jurisdictions are turning versions of previously voluntary self-governance mechanisms into regulatory requirements, pilot development of modular (multistakeholder, co-regulatory governance) standards and mechanisms to be used across multiple regulations, companies, and countries, providing the specificity many regulations lack on exactly how to implement intended rules around transparency, data access, and other requirements. This could also help ease pressures to turn these practices into compliance exercises in their entirety.

Invest in Cross-Platform Research Focused on Ecosystems and Incentives

- 5.4 Develop archetypes of problematic actors to support the identification of effective T&S levers, informed by their respective motivations and incentives.
- 5.5 Develop frameworks and tools to support the mapping of abusive actors and their cross-platform presences. Particularly as new mediums and technologies emerge, approaches to content mapping and response will shift. Abusive or problematic actors, as well as the incentives structures (such as a particular monetization model) driving them, will be constants even as content takes different forms. Increased focus on mapping and understanding actors and their drivers will equip sectors working across T&S to develop products that are safer by design, and to prevent and mitigate the abuse of new digital surfaces or tools.

²⁰ For a longer list of recommendations regarding open-source tooling, see [Annex 2: Building Open Trust and Safety Tools](#).

INVEST IN INDEPENDENT RESEARCH TO ADDRESS CRITICAL GAPS IN KNOWLEDGE

The following are gaps in knowledge that require urgent attention, on topics certain to have catalyzing impact on the future web. These focus on better understanding trust and safety as a discipline, and the three areas of tech innovation identified in the full report driving the direction of the future web. This includes generative AI, decentralization, and experiential, immersive, and augmented technologies.

Generative AI (GAI)

- 6.1** How might GAI be applied to a range of content-moderation challenges including the quality of classifiers in non-majority languages, reducing human exposure to harmful content, detecting influence operations, or other legacy issues?
- 6.2** How might GAI be leveraged to supercharge existing harms or create new harms—challenging existing mechanisms and processes for trust and safety?
- 6.3** How effective might technological innovations like watermarks and other digital provenance techniques be in enabling society to adapt to GAI innovations with more meaningful informed consent and awareness?
- 6.4** Could model training and the development of “pre-cleaned” foundation-model training datasets be used to de-bias and de-toxify large-language models (LLMs), and to strip personal information from them as well?²¹

Trust and Safety as a Discipline

- 6.5** What are the mental-health and psychosocial impacts of different kinds of online spaces (ensuring diversity in the communities researched), comparing social media, gaming platforms, and XR?
- 6.6** What are the longitudinal impacts of common T&S tooling approaches? This might apply to specific things—like parental screening tools, or mechanisms to flag and review viral content—but should be focused on enabling continued iteration and the development of evidenced-based policies and standards.
- 6.7** What are the most common risks elevated by different business models, and how can they be mitigated?
- 6.8** What are the impacts of different platforms on highly marginalized communities across a range of geographies and cultures? Particular attention should be given to law enforcement and other state-affiliated actors’ use of platforms to conduct surveillance, spread harmful narratives targeting minority groups, or persecute political activity.
- 6.9** What limitations within current funding models in philanthropy and foreign assistance currently undermine civil-society capacity to maximize its role within the broader T&S ecosystem, particularly given the central role played by state actors in T&S and growing crackdowns on funding pathways for organizations and independent researchers. How could new models be developed that respond to this growing geopolitical trend?

²¹ For a longer list of recommendations, see [Annex 6: Learning from Cybersecurity, Preparing for Generative AI](#).

Decentralization

- 6.10** What are the risks and challenges posed by disinformation and manipulative behavior on federated platforms, and how do these risks differ from those created on centralized social media services?
- 6.11** What are existing moderation capabilities—technical and otherwise built into federated services—and how effective are they at addressing behavioral and scaled threats? What capabilities do we need to build for the future?
- 6.12** What are appropriate governance frameworks and organizational structures for this work in a decentralized context? Are there good examples of norms adopted by these communities?
- 6.13** What can be learned from community-moderated systems, and how can that be applied more broadly? How can lessons from Wikimedia, Reddit, gaming messaging boards, and other forums apply to decentralized contexts?
- 6.14** Are there viable business models based on decentralized architectures like blockchain, and are there financial innovations built upon them? Where are they most likely to succeed and fail?²²

Extended Reality (XR), Metaversal Technologies Immersive, Experiential, and Augmented Technologies

- 6.15** What are promising content-moderation models and technologies that are privacy respecting and scalable for real-time, audio, and experiential contexts?
- 6.16** Does interacting with immersive technologies have any unique or amplified impact as compared to other digital technologies? How does it compare to in-person interactions?
- 6.17** What are the physical- and mental-health impacts of leading experiential technologies?
- 6.18** Where do existing regulatory frameworks need updating for digital, immersive applications, and what are their limits and gaps? For example, is there a scenario in which the biometric data collected through wearables and virtual-reality devices would be treated as sensitive and protected health-related data? Do non-tech-related regulations and rules apply? If so, when?
- 6.19** Are there unique risks related to potential advertising on the basis of biometric data?
- 6.20** What are the regulatory layers at play as gaming, interactive, and information-related platforms increasingly intertwine?

²² For a longer list of recommendations, see [Annex 6: Learning from Cybersecurity, Preparing for Generative AI](#).

CONCLUSION

This executive report captures the key findings of the Task Force for a Trustworthy Future Web as well as its recommendations for specific, actionable interventions that could help to overcome systems gaps the task force identified. Ideally, this report should be read within the broader context of the task force's comprehensive report, *Scaling Trust on the Web*, which offers deeper insights into the questions that were most carefully considered and addressed, as well as more granular recommendations for future work.

The accompanying annexes provide, respectively:

- 1 A review of how the current T&S field has emerged, the knowledge and practices that have been developed within it, and where it offers opportunity as well as requires evolution and advancement.
- 2 An analysis of where tooling necessary for T&S might benefit from intentional and collective investment and focus.
- 3 An examination of the role that children's rights and inclusionary participation models can play in debates regarding child safety online.
- 4 An introduction to the gaming industry, highlighting its influence on online spaces now and in the future.
- 5 An assessment of the T&S capabilities of federated platforms, with a particular focus on their ability to address risks like coordinated manipulation and disinformation.
- 6 A review of lessons that could be learned from the evolution of the cybersecurity industry, as well as a forecast of how generative AI may impact T&S.

We are on the precipice of a new digital era. It is our hope that the insights captured in *Scaling Trust on the Web* galvanize investments in systems-level solutions that reflect the expanding communities dedicated to protecting trust and safety on the web, the trailblazers envisioning the next frontier of digital tools and systems, and the rights holders whose futures are at stake.

ACKNOWLEDGMENTS

The Task Force for a Trustworthy Future Web launched in February 2022, bringing together more than forty experts in policy, AI, trust and safety, advertising, gaming, civil rights, human rights, law, virtual reality, children’s rights, encryption, information security, community organizing, product design, digital currency, Web3, national security, philanthropy, foreign assistance, and foreign affairs.

Over a five-month sprint, through interviews, expert roundtables, thematic discussions, document reviews, and briefings, task force members shared hard won lessons about what has worked and what hasn’t worked over twenty years of striving to build safe, useful spaces where humans can come together online. **This sprint had four goals:**

- 1 Map systems-level dynamics and gaps that will continue to impact the trustworthiness and usefulness of online spaces regardless of technological change.
- 2 Highlight where existing approaches will not adequately meet future needs, particularly given the emergence of new “metaversal” and GAI technologies and the diversification of online spaces.
- 3 Identify significant points of consensus across the membership’s broad range of perspectives and expertise.
- 4 Generate concrete recommendations for immediate interventions that could fill systems-level gaps and catalyze safer, more trustworthy online spaces, now and in the future.

Scaling Trust on the Web captures the task force’s key findings. It provides a brief overview of the truths, trends, risks, and opportunities that task force members believe will influence the building of online spaces in the immediate, near, and medium term. It summarizes recommendations identified throughout the task force’s work for specific, actionable interventions that could help to overcome systems gaps the task force identified. Its six annexes provide deeper insights into the questions that were most carefully considered and addressed.

All task force convenings and interviews were conducted under the Chatham House Rule and any quotes included in this report are with permission. **The analysis reflected in *Scaling Trust on the Web* does not represent the individual opinion of any member of the task force or any contributing organization to the task force. Rather, it serves to consolidate collective research, feedback, and contributions gathered over a five-month period.** The task force staff has striven to reflect feedback fairly, accurately, and thoughtfully, but any errors or omissions are our own. Each annex was drafted through a unique methodology that is enumerated at the end of the annex.

DFRLab is indebted to the task force’s members, contributing expert organizations, and contributing experts for the time, care, candor, creativity, wisdom, and overall esprit de corps they gave to this fast-paced and iterative endeavor. Each contributor volunteered their time over an extraordinarily busy five months.

Because the task force was designed as a sprint, rapid pivots and tight review deadlines were the norm. The task force director would like to extend her most sincere and personal thanks to all our contributors for their considerable graciousness, flexibility, and trust as the task force’s work evolved. In addition, for their superlative support, guidance, and diligence, she would like to express enormous gratitude to Nikta Khani, associate director of the task force; Rose Jackson, Eric Baker, and Graham Brookie of DFRLab; and MaryKate Alyward of the Atlantic Council. She would also like to thank her husband, Evan Handy, and her children for their many contributions behind the scenes to the success of this endeavor.

The Task Force for a Trustworthy Future Web was generously supported by [Schmidt Futures](#) and the [William and Flora Hewlett Foundation](#), and DFRLab would specifically like to thank Eli Sugarman for his unflagging commitment, creativity, trust, and good humor throughout the development and execution of this initiative.

TASK FORCE MEMBERS

An asterisk denotes a task force member who also served on the steering committee, collaborating from the outset to inform the scope and objective of the task force's work. Steering committee members graciously took on extra responsibilities such as reviewing materials in advance of the broader membership or participating in prebriefings before task force calls.

Chinmayi Arun
Executive Director
Information Society Project,
Yale Law School

Savannah Badalich
Senior Director of Policy
Discord

Tami Bhaumik
VP of Civility and Partnerships
Roblox

Lauren Buitta
CEO
Girl Security

Agustina del Campos
Director
Center for Studies on
Freedom of Expression

John Carlin
Partner
Paul Weiss

Daniel Castaño
Law Professor
Universidad Externado
de Colombia

Dr. Rumman Chowdhury
Responsible AI Fellow
Berkman Klein Center,
Harvard University School

Nighat Dad*
Executive Director
Digital Rights Foundation

Michael Daniel
CEO
Cyber Threat Alliance

Justin Davis
CEO
Spectrum Labs

Emma Day
Human Rights Lawyer
Nonresident Fellow
DFRLab

Dante Disparte
Chief Strategy Officer
Head of Global Policy
Circle

Kat Duffy*
Director
Task Force for a
Trustworthy Future Web

Alex Feerst*
CEO
Murmuration Labs
Cofounder
Digital Trust &
Safety Partnership

Camille Francois*
Global Director of
Trust and Safety
Niantic

Grace Githaiga
CEO
KICTANet

Inbal Goldberger
VP of Trust and Safety
ActiveFence

Brittan Heller*
Affiliate
Stanford Cyber Policy Center
Nonresident Senior Fellow
DFRLab

Sue Hendrickson
Executive Director
Berkman Klein Center,
Harvard University

Rose Jackson*
Director
Democracy + Tech Initiative
DFRLab

Lea Kissner
Former Chief Information
Security Officer
Twitter

Bertram Lee Jr
Senior Policy Counsel
Data, Decision-Making
Artificial Intelligence
Future of Privacy Forum

Jade Magnus Ogunnaiké
Vice President of Campaigns
Color of Change

Katherine Maher
Former CEO and
Executive Director
Wikimedia Foundation

Mike Masnick
Founder
Techdirt and Copia Institute

John Montgomery
VP
Global Brand Safety
Group M

Sidney Olinyk*
CEO
Duco Experts

Riana Pfefferkorn
Research Scholar
Stanford Internet Observatory

Leah Plunkett
Meyer Research
Lecturer on Law
Harvard Law School
Faculty Associate
Berkman Klein Center for
Internet and Society,
Harvard University

Victoire Rio
Digital Rights Advocate

Yoel Roth
Technology Policy Fellow
UC Berkeley Goldman School
of Public Policy

Eli Sugarman*
Fellow
Schmidt Futures
Talent Ventures

Dr. Kimberly Voll
Cofounder
Fair Play Alliance

Tiffany Xingyu Wang
Chief Marketing Officer
OpenWeb

Timoni West
Vice President and
General Manager
Unity

Maya Wiley
President and CEO
The Leadership Conference
on Civil and Human Rights
The Leadership Conference
Education Fund

Charlotte Willner*
Executive Director
Trust & Safety Professional
Association (TSPA)

Dave Willner
Head of Trust & Safety
OpenAI

Nicole Wong
Former Deputy Chief
Technology Officer of the
United States

*Steering Committee

EXPERT CONTRIBUTING ORGANIZATIONS

[ActiveFence](#)

[All Tech Is Human](#)

[Berkman Klein Center](#)

[Color of Change](#)

[Digital Trust & Safety Partnership](#)

[Duco Experts](#)

[Fair Play Alliance](#)

[Forum for Democracy and Information](#)

[Global Network Initiative](#)

[Integrity Institute](#)

[Leadership Conference on Civil and Human Rights](#)

[Spectrum Labs](#)

[Tremau](#)

[Trust & Safety Professional Association](#)

[WITNESS](#)

CONTRIBUTING EXPERTS

The task force benefited from initial foundational analyses by Duco Experts, 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. Additionally, we thank the following individuals for their contributions to task force coverings and/or written products.

Meri Baghdasaryan

Expert

Duco Experts

Ted Han

Director of Mozilla Rally

Mozilla

Betsy Masiello

Cofounder

Proteus Strategies

Safa Shahwan Edward

Deputy Director

Cyber Statecraft Initiative
DFRLab

Georgia Bullen

Executive Director

Superbloom

Katie Harbath

CEO

Anchor Change

Angela McKay

Director of Research

& Partnerships

Trust & Safety
Google

Derek Slater

Cofounder

Proteus Strategies

Eline Chivot

Senior Adviser on Digital

Policy and Economic Affairs

European People's Party

Trey Herr

Director

Cyber Statecraft Initiative

DFRLab

Sarah Oh

Cofounder

T2

Matthew Soeth

Head of Trust and Safety

Spectrum Labs

Anne Collier

Founder

Net Safety Collaborative

Julie Hollek

Director of Data Science

Mozilla

June Park

Asia Fellow

International Strategy Forum

Alex Stamos

Director

Stanford Internet Observatory

Olivia Conti

Community Health

Twitch

Sara Ittelson

Partner

Accel

John Perrino

Policy Analyst

Stanford Internet Observatory

David Sullivan

Executive Director

Digital Trust & Safety
Partnership

Eric Davis

Trust & Safety

Security

Privacy Consultant

Jeff Jarvis

Director

Tow-Knight Center for
Entrepreneurial Journalism
The City University of
New York

Lauren Quittman

Tech Policy Manager

Duco Experts

Lauren Wagner

Fellow

Berggruen Institute

Renee DiResta

Technical Research Manager

Stanford Internet Observatory

Jen King

Data Privacy and Policy Fellow

Stanford Institute for
Human-Centered AI

Ashwin Ramaswami

Researcher

Plaintext Group

Sarah Williams

Senior Manager

Systems & Tools
Pinterest

Brian Fishman

Cofounder

Cinder

Hilary Ross

Affiliate

Berkman Klein Center,
Harvard University

Dr. Sara M. Grimes

Associate Professor

Faculty of Information
University of Toronto

Jaz-Michael King

Program Lead

IFTAS

ANNEX 1

SCALING TRUST ON THE WEB


CURRENT STATE OF TRUST AND SAFETY

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

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 Klionicles*; 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.

ANNEX 2

SCALING TRUST ON THE WEB


BUILDING OPEN TRUST AND SAFETY TOOLS

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

ANNEX 2**BUILDING OPEN TRUST AND SAFETY TOOLS****TABLE OF CONTENTS**

Introduction	2
Methodology	3
The T&S Tech-Tooling Landscape	4
How Tools Are Built Today	4
Building Shared Technology Tooling	6
Detection	6
Hash Matching: Detecting Exact or Near-Exact Matches of Violative Content	7
Classifiers and Other Automated Assessment of Previously Unseen Content	8
Enforcement	10
Rules Engines	10
Queueing and Workflow-Management Tools	10
Measurement	11
Transparency	11
Building a Hub for Best Practices	12
Looking over the Horizon	12
Artificial Intelligence	12
Evolving Service Types	13
eXtended Reality (XR) and the Metaverse Technologies	13
Decentralization of Services and User Controls	14
Age Assurance	15
Authorship & Acknowledgments	16
APPENDIX Open and Shared T&S Tools References	17

INTRODUCTION



Trust and safety practices (T&S) are often misperceived as only a policy challenge for tech services to tackle, a matter of whether they have the right policies in place for managing potentially harmful content and behavior to keep users safe.¹ In reality, there is a technical implementation layer that is highly complex and is often built over time as a homegrown tooling suite and organizational structure as a service begins receiving user complaints about abuse on the platform. T&S is as much a logistics challenge as a policy challenge—a matter of facilitating effective decision-making about content and conduct, undergirded by technology.

While T&S has essentially existed as long as Internet services have, it is still maturing as a field. In recent years, organizations have begun to fill crucial gaps—for instance, by providing training and support to practitioners in developing policies and organizational structures, as well as support for organizations in developing risk-assessment frameworks.

T&S tooling is an area that remains ripe for intentional, collective investment and focus. More effective, openly available tooling—as well as more accessible best practices for development of T&S tools—could lower barriers to the development of, and competition among, a diversity of services, making it so that each organization does not need to reinvent the wheel. Moreover, it can help address what is essentially a market failure—individual services may not internalize all the social costs of harmful content and behavior, and, thus, may not invest sufficiently in socially optimal T&S.

In this paper, we consider the role open tools do, and could, play in improving T&S for a broad array of stakeholders, and where philanthropic (or other) investment might most usefully be directed. By open, we mean to be inclusive of: open-source software tools; open or pooled data; and shared-tooling solutions that may rely on proprietary software, technical resources, or services, but which are deliberately structured to be available to a wide range of platforms.

¹ We use the term trust and safety (T&S) throughout to refer to the field and practice of determining appropriate content and behavior on an online service, and managing content and conduct-related risks. The teams that carry out this work go by a variety of names; for instance, they may also be referred to as “integrity” teams. Furthermore, T&S will include product and engineering teams themselves and will intersect with other product and engineering teams in other departments. For simplicity, we simply use the umbrella term T&S.

There are opportunities at each layer of the T&S operation to support the ecosystem through open tools. Our key conclusions are as follows.

- 1** The logistical aspects of T&S operations appear ripe for the development of robust open tooling. Services depend on a range of tools to manage key workflows, including confirming enforcement decisions, logging and measuring decisions over time, and ensuring transparency to the public, policymakers, and others. Relevant tools include rules engines, which automate the processing and management of potentially violative content or behavior; review consoles, which provide interfaces for human review; and case-management systems, which allow services to track actions taken on individual instances of detected content and behavior.
- 2** Some types of detection tools provide clear opportunities for open solutions. Services use hash-matching tools to detect exact and near-exact matches of previously identified content, and some of these tools are already open source. In deploying these tools, services must consider not only their effectiveness, but also the extent to which public access to a tool's inner workings may also benefit bad actors who want to circumvent detection. Services can also benefit from best practices and toolkits that help them build classifiers that can help assess new, previously unseen content or behavior.
- 3** Content-specific detection tools present a complex challenge, demanding greater forethought to overall governance and institutional support. While a wide array of services may have policies against common types of content (e.g., hate speech), services' individual policies vary, no one tool will suit all, and detection tools must be updated over time. Moreover, these tools may raise complex legal questions—for instance, those related to processing of personal data. In turn, creating shared databases of violative content or content-specific classifiers raises many questions beyond simply technological design. While this is a more complex endeavor, it can provide significant utility.

More generally, ongoing maintenance, improvement, and other stewardship can shape the utility of open tools. This is particularly the case for T&S, in which tooling needs to be responsive to an ever-changing adversarial landscape. While open tools exist that tackle some of the opportunities above, they are inherently limited by the dispersed and disaggregated nature of tooling, as well as the lack of support for ongoing development. In turn, there is an opportunity to build an institutional hub for best practices that could provide technical tooling support for T&S. Such an institution could serve as a one-stop shop for practitioners who are navigating the tooling landscape, as well as contribute to development of new tools, and—critically—act as a steward of open-source tools that other actors are willing to contribute but unable to maintain in an ongoing way.

METHODOLOGY

This paper builds on workshops and interviews with about fifty T&S expert professionals, who could speak to the needs of a variety of types of organizations (companies, nonprofits, vendors) of different sizes and focuses. While this is only an initial sample of perspectives, the report synthesizes common themes to suggest a path forward that could benefit the ecosystem. We held five workshops, one of which was in person and four of which were virtual, with practitioners, experts, and stakeholders in T&S tooling. Participants included T&S practitioners from a wide variety of services, including some of the largest platforms and many smaller service providers, cutting across many different use cases (e.g., social media, commerce, real-time communications, dating, media sharing, discussion forums) and media types. We also consulted specifically with noncommercial service providers, vendors building enterprise solutions specific to T&S challenges, and academics. Following these workshops, we conducted several independent interviews with experts, and

compiled desk research and data made available to us by GitHub to produce the analysis in this report. While the workshops and interview sessions were conducted under the Chatham House Rule, we have included a few quotes with permission.

THE T&S TECH-TOOLING LANDSCAPE

In this report we are focused on technologies that support T&S operations, especially as a service starts to develop and scale. Broadly speaking, T&S operations can be thought of as an iterative loop moving through four distinct phases, and then back again: detection, enforcement, measurement, and transparency.² While this can be a linear process, it is often more interactive, as systems must not only be improved over time but also respond in real time to actors who try to game or subvert the system.

DETECTION

Organizations identify potentially violative content or conduct based on reports received from users, or from automated tools. These tools either match a piece of content against a database of known violations or rely on advanced technology to assess the likelihood that a piece of content or given behavior is violative.

CONFIRMATION AND ENFORCEMENT

Once potentially violative content or conduct is detected, organizations must decide what action to take—for example, removing content, downranking content, or banning users. Automated rules engines help manage this process, taking immediate action in response to detection in certain cases (e.g., where automated tools are highly likely to be accurate and the potential harm is severe), or routing and prioritizing content or conduct for manual review. Software tools are also relied on for case and workflow management, providing T&S teams with an interface to track and take action.

MEASUREMENT AND INFRASTRUCTURE

Tools are used to log data about enforcement decisions and subsequently analyze them for the purposes of measuring enforcement effectiveness, audit and detection of abuse of the system, and further training of automated detection mechanisms, among other things. These data are also analyzed for internal management and regulatory compliance.

TRANSPARENCY

Organizations need tooling to message enforcement decisions back to users, facilitate appeals processes, and report on T&S operations to the public, policymakers, and other audiences.

² While we have used this simplified distillation, one might also think of T&S in terms of a “tech stack.” A tech stack is a set of tools that serves particular purposes and is aligned to a product-development process, which can broadly be generalized to backend, midlayer, and frontend components. See e.g., Zoom’s discussion of its T&S “tech stack.” From this vantage point, detection, confirmation and enforcement, measurement, and transparency are the relevant goals of the “stack.”

In addition to this operational loop, it is worth noting the integral role that product development as a whole can play. Product features can facilitate addressing harms, and can also proactively promote content and behavior that are understood to be additive to the community. For instance, services may develop methods for detecting and surfacing high-quality news content, or use feedback mechanisms to inform detection of other types of content. Moreover, product features may include specific controls or tools for users to manage or address harmful content or conduct. For instance, users can block or mute messages from particular users, or users might select filters that screen out certain content for themselves (while the content remains available to others on the service). These features may produce data that are then used to inform how T&S detection tools, or other parts of T&S operations, function. This broader suite of product features is not the focus of this report, although we will come back to user tools briefly in the final section.

On certain services, community-moderation systems can play a central role. These systems allow users to moderate comments, posts, and other activities within a service. They figuratively sit atop an organization's T&S operations. Examples include Discord and Reddit's models for moderating content in different channels; and Facebook and Nextdoor's models for moderating user behavior in groups. In the case of Wikipedia, effectively all content editing and moderation is managed by users themselves, and the [Wikimedia Foundation's T&S function](#) focuses on supporting the community and a specific range of abuse types (e.g., legal removal requests). To some extent, users who fill community-moderation functions face similar issues to a service provider's T&S teams at the platforms themselves: detecting violative content and making enforcement decisions. At the same time, tools made available to these moderators will also vary, and they are not our key focus here.

In evaluating opportunities for open and shared tooling solutions, we have focused on elements of the stack that are content agnostic and elements that are content specific. For example, a workflow or case-management tool for a given type of media (e.g., text, audio) may need to incorporate categories for hate speech, harassment, and many other content categories, but the most basic workflow elements are a universal part of every T&S operation in every organization and are necessary regardless of the specific category at issue. In contrast, a classifier to detect content that violates a specific rule (e.g., hate speech) will be specific to the type of content being detected.

The degree of content specificity has implications for how open a tool can ultimately be. Content-agnostic tooling is generally going to be more amenable to open development and shared solutions, for the simple reason that it is more universally applicable. While there are also a number of opportunities to open up content-specific tooling and solutions, these may face more limitations and challenges given the need for greater customization by the organizations that implement them.

HOW TOOLS ARE BUILT TODAY

Historically, when organizations had needs for T&S tooling, the build-or-buy decision was made for them by a paucity of fit-for-purpose tooling solutions on the market. In addition, services must carefully consider the privacy and security demands of T&S; for example, T&S may involve review of both public and private content and behavior, and services must carefully delineate who can access both T&S tools and resulting logs. As a result, custom solutions have been developed in house time and time again, often solving variations on a more generalized problem seen before by other services.

In recent years, a number of vendors—some start-ups, some larger enterprises—have emerged to provide tooling solutions to T&S problems. While these companies signal the emergence of market solutions for tooling needs, T&S professionals who we interviewed voiced concerns about dependence on a limited set of firms, particularly as start-up vendors have tended to eventually be acquired and brought in house by larger companies. For instance, in 2018, Twitter acquired anti-abuse provider Smyte, and the tool was taken

off the market. Our initial scan of market data suggests that as many as one-third of all T&S start-ups in the past decade were acquired by other companies. In many of these cases, the former customers of the start-up may lose access to an external tool on which they had relied, and must then reinvent the wheel with in-house development or by finding another vendor.

Open tools exist today, and are put to use to varying degrees. In our interviews, practitioners suggested that open tools might provide a starting point from which service providers can build, but that there is not a go-to set of tools on which many could readily depend. Instead, there is a vast array of different options available, and it is challenging for services to determine the suitability of any given tool.

A core challenge in providing open-source tools is the need for ongoing maintenance and customization. Nothing in T&S is “set and forget”—tools need constant maintenance to stay relevant to the evolving threat landscape online. Practitioners with whom we spoke pointed to the risk of relying on an open-source solution if it is not well supported, or if there is a chance the originating organization may cut support in the future. As we will discuss further below, these challenges with the open-source tooling ecosystem today present an opportunity for focused investment and support.

BUILDING SHARED TECHNOLOGY TOOLING

In the sections that follow, we describe the elements of the T&S operational loop, and the varying degree to which these elements are amenable to open-source tooling and shared solutions. Each operational element has different component parts, some of which are content specific and others of which are content agnostic.

DETECTION

The most basic mechanism through which platforms detect potentially harmful content and behavior is a user-initiated action; most platforms provide some mechanism, often a flag or similar button to “report a violation,” for users to report violative content or harmful behavior. While this functionality is common, it ultimately needs to be tailored and customized to fit a service. Best practices and reference implementations could still usefully inform how services build their own systems, and, as we discuss later, ensuring these user-initiated actions are routed to the appropriate teams is an element of workflow management in which open tooling could be helpful.

Similarly, some services use “trusted partner” or “trusted flagger” mechanisms that allow nongovernmental organizations (NGOs), governments, and other organizations to access more sophisticated reporting tools than are available to the average user. These programs often route trusted partners’ reports with different prioritization, as organizations rely on partners for language and cultural or other expertise in evaluating content. On the one hand, these tools are also customized to integrate into a given product or service, and, thus, are not prime candidates for open tool building. On the other hand, these programs are increasingly becoming required of some platforms by regulation, and may, therefore, become standardized to varying degrees. What is more, participants in our interviews noted that these systems were often not tailored for a diversity of partners, including organizations from the Majority World.

In addition to detecting abuse reactively via user-initiated action, services also deploy automated, proactive detection tools for violative content and behavior. The scale of these efforts can vary with the purpose and scale of the organization, but they generally exist at least in some form (for example, to address spam).

In particular, we spoke with experts about two common tools: hash matching for exact and near-exact content matches; and classifiers built with machine learning or other AI tools to assess new, previously unseen content for potential violations.

HASH MATCHING: DETECTING EXACT OR NEAR-EXACT MATCHES OF VIOLATIVE CONTENT

Hash matching is a method used to detect content that is exactly the same or nearly the same as previously identified content. Exact hash matching identifies exact matches to known violative content. “Fuzzy matches” and perceptual hashes are used to find content that is nearly identical to previously identified content, but different enough to be missed by exact hash matching.³

Fundamentally, hash matching is a content-agnostic method for creating and identifying hashes of content and matching that content to known violations, and both exact and “fuzzy match” methods can be, and have been, open source. For instance, common hash functions are [readily available as open-source libraries](#). Companies have also contributed code to improve methods. For instance, Facebook has opened its photo and video hash-matching tools, [PDQ](#) and [TMK+PDQF](#), respectively, as well as a more comprehensive [Hasher-Matcher-Actioner tool](#) that facilitates labeling, matching, and actioning violative content.

Hash functions are designed to impede reconstructing the original content simply from the resulting hash. While some argue that open tools may be more vulnerable to adversarial attacks, others suggest that “security through obscurity” is not effective here, and that the [benefits](#) of open-source contributors overall support the effectiveness of these tools.

Of course, effective hash matching depends on having a database of hashes. Typically, a service maintains a database of hashes based on content it has already addressed. In addition, services might rely on shared databases of hashes, matching content on their service against a database of previously identified content from elsewhere.

The most common instance of this approach exists with child sexual-abuse material (CSAM). For instance, the National Center for Missing and Exploited Children (NCMEC), International Watch Foundation (IWF), and other similar organizations maintain databases of CSAM (pursuant to legal guidelines), and services can then check against these databases to take action against violative content on their systems.⁴ In many cases, services access shared databases by using hash-matching tools that are offered as a centralized, though broadly available, service in the context of addressing CSAM. [Microsoft](#), [Google](#), and [Cloudflare](#) have all built CSAM hash-matching engines that check content against databases of known CSAM, and which they make available to other platforms via API or, as in Cloudflare’s case, customers.

In recent years, databases of hashes have also been developed for shared use in other contexts. Most notably, the [Global Internet Forum to Counter Terrorism \(GIFCT\)](#) has developed a [cross-platform, shared hash database of terrorist and violent extremist content \(TVEC\)](#) that member organizations can use to identify content on their platforms. Unlike CSAM, TVEC does not have a universal definition and is not universally illegal, and each organization can decide to act on this content in different ways.

A similar institutional solution has been developed for nonconsensual intimate-image abuse. [StopNCII.org](#) is a hash database operated by the UK Revenue Porn Helpline, a nonprofit organization in the United Kingdom (UK). Adults are able to generate a hash of intimate imagery that was created without their consent; this hash is created locally on a user’s device, and then only the hash is sent to StopNCII.org for inclusion in the database. Participating companies will run their content against this database to detect and remove matches.

² Here, we use hashing in the way the Office of Communications (OfCom) does in its extensive report: “Hashing is an umbrella term for techniques to create fingerprints of files on a computer system.” See also Hany Farid’s [“An Overview of Perceptual Hashing.”](#)

³ For a deeper analysis of this topic, see [Annex 3: Respecting Children as Rights Holders](#).

These content-specific hash databases, and associated hash-matching tools, are another area for possible investment, but they raise much more complexity. Investment in additional hash databases and matching solutions needs to be regarded as an institutional challenge—just as much as, if not more than, a tooling challenge—engendering the trust of a wide range of stakeholders.

On the one hand, practitioners noted the utility of such tools, particularly in the case of CSAM, where the harm is severe and the content is universally illegal regardless of context.⁵ On the other hand, the utility of these databases depends on strong, trustworthy governance regarding their contents. Concerns about bad actors using the database to reverse engineer and access harmful content, or accessing the hash function to circumvent it, may encumber the full openness of available tools. There is also a risk that data will be improperly included in the database, and, as a result, that organizations will be overly restrictive in removing content.

A related concern with investing in developing broadly shared databases is they would facilitate the development of “[content cartels](#),” even if done unintentionally. Because these systems can be expensive to develop and maintain, the concern is that the entire ecosystem would default to the easiest and most available, creating a de facto standard.⁶

Along with detecting violative content, another detection challenge that many T&S professionals cited is the need to share intelligence across platforms and receive intelligence from trusted partners. This is particularly the case for detection of behavioral patterns that contribute to harmful content online. For example, identifying and enforcing against disinformation often requires identifying clusters of fake accounts that are found to be acting in concert, or other forms of coordinated inauthentic behavior. Today, some services have developed arrangements to share information with industry peers, but doing this in a way that respects privacy is challenging. Regardless, creating an institution to facilitate threat sharing would raise many of the complexities of the hash databases noted above. Again, the challenge here is as much institutional as it is technological.

CLASSIFIERS AND OTHER AUTOMATED ASSESSMENT OF PREVIOUSLY UNSEEN CONTENT

Hashing solutions help with identifying exact or near-exact matches to previously seen content. But what about new content? Services use a variety of approaches, including automated systems that monitor for the characteristics of bots used to manipulate platforms; text-analysis tools; and a variety of approaches based around machine learning and artificial intelligence (AI). For instance, through machine learning and other AI techniques, services create and deploy “classifiers” that automatically assess new content and behavior to assign scores that reflect the likelihood of violations. These tools are a linchpin tool for detection of abuse, automating a first-pass evaluation before passing things into a queue for human review.

Practitioners noted that the T&S ecosystem would benefit from best practices and toolkits that facilitate the development and evaluation of classifiers and other detection tools. Services could benefit from guides and reference implementations that assist in the process, and approaches for measuring and evaluating efficacy. For example, Google has provided a [reference implementation](#) of how to use the open-source TensorFlow platform for creating content-moderation tools. While the implementation uses an example classifier created for detection of toxic content, anyone can take this model, install it, and run it on their own dataset to develop a classifier that is fit for purpose.

⁵ In addition, in the context of CSAM, there are restrictions on even possessing the content. As such, there are some benefits to relying on a third party that has the requisite permission to operate a database of this sort.

⁶ In addition, to the extent the tools rely on passing a service’s content to a third party for analysis and matching, it raises potential competitive and privacy concerns for the originating provider. There may be ways to address this concern by creating a hash at the client side and then passing that to the server. For instance, Microsoft has begun trialing such a system for its PhotoDNA system. See: Microsoft Moderator Service API Documentation, “[Match Edge Hash](#).”

Building open, content-specific classifiers is also possible, but raises institutional challenges as much as technological ones. A wide variety of open classifier tools already exist.

- ▶ An initial analysis of open-source code repositories on Github, based on only a few dozen keyword topics, found more than five hundred libraries related to content classifications; a robust analysis would surely find multiples more.
- ▶ [Hugging Face](#) has cultivated a developer community building open-AI models, and it features datasets and detection tools for different types of content.
- ▶ Social media platform Bumble released a tool for “lewd photo detection.”
- ▶ Jigsaw, an Alphabet company, has contributed its [conversational-AI-moderator code](#) to Github, which can be used to detect toxic comments; Jigsaw has also released this as [Perspective API](#) that others can use.
- ▶ Startups like [Unitary](#) are also contributing classifiers to the market, and are committed to building more with open source.

Despite the appearance of a robust set of open classifiers, practitioners suggested that, today, these tools are of only limited utility. To begin with, it is challenging to even know what is available—there is no trusted source or compilation of all the open tools that exist. Even when organizations have deployed classifiers that were originally developed externally—in an open-source fashion, or in cases where classifiers have been brought in via acquisition of start-ups—T&S professionals must heavily customize the tools. Even with a tool available as open code, it can be difficult to evaluate and compare its performance, including how it was trained. For example, we heard in our research that even with a seemingly widely used classifier, such as [Yahoo!’s Open NSFW tool](#) (which focuses solely on detecting pornography), many organizations that use it take it as a baseline input on top of which they build further customization for their platforms’ specific needs. Expert Adelin Cai, who is a co-founder of the Trust & Safety Professional Association and worked on T&S teams across a number of companies, has seen multiple instances of services paying for third parties to perform initial screening of content on the platforms, as well as to provide lists of keywords related to potentially harmful content. Nevertheless, the services still need to do in-house customization due to the uniqueness in how each organization would use the output, so “wouldn’t it be great if there were open options for those organizations that just need somewhere to start.”

Classifiers also need active stewardship to be effective. Like spam filters or other systems intended to identify or screen certain types of content, classifiers operate in an environment where adversaries are continuously working to game the system and push their violative content through. As a result, any investment in this tooling needs to account for ongoing governance of the tool, ideally embedded inside an organization tasked with its ongoing upkeep.

Moreover, deployment of third-party classifiers raises issues similar to those of the hash databases discussed above. While access to code can help support a greater degree of transparency and accountability, it is insufficient to fully understand how the classifier operates. To deploy a classifier, a service provider has to trust the data and that the training of the tool was done in a way that aligns with its values and policies—and that may not be the case. In every case, classifiers may carry with them bias, unfairness, or inattention to a variety of other contextual factors. Even something as seemingly mundane as a profanity classifier could be trained in a way that is more or less attuned to certain vulnerable communities.

Nevertheless, participants repeatedly pointed to the limited geographic reach of existing classifiers as a key market failure, and an area in which investment in open and shared classifiers could address gaps in a meaningful way. This may be the case with languages and communities in the Majority World, for example. Data availability in certain languages may be highly limited, and practitioners noted that companies whose

primary revenue-driving markets are English language and culturally Western are unlikely to invest in building high-quality classifiers for other markets and languages.

Another opportunity for increased investment is the contribution of datasets on which classifiers can be trained and evaluated. Particularly for small organizations, building the underlying dataset can be a lot of work, as it requires both the underlying content and the human and computational investment in labeling the dataset. While a user of these sets must still be discerning about the underlying content before using it as an input in the model, these could still provide useful starting points.

ENFORCEMENT

After detection of harmful content, T&S teams must confirm an assessment of the content and implement an enforcement decision. Many tools are currently used to do this both manually and automatically, and some of them might be good candidates for open and shared development. In this area, we found relatively little when it comes to existing, open tools tailored for T&S; however, practitioners suggested that these aspects held promise for shared solutions.

RULES ENGINES

As the number of content flags scales with a platform's growth, many T&S teams quickly find themselves undated beyond what they can manually react to in a timely fashion. Rules engines are built to provide a first run of automated processing on high volumes of content, and several experts spoke of them as critical tools in the T&S toolkit. These engines automate some enforcement decisions, but primarily route and prioritize decisions for content and conduct that has been classified already.

While different services will create different rules, rules engines themselves are a general need, and are a potential opportunity for building open tools. Alex Feerst, a leading expert and former general counsel and head of trust and safety at Medium, was among the many practitioners to note how rules engines were a critical part of the “core plumbing” of T&S operations:

“Every company needs to think about the logistics of detecting and reviewing content and conduct. They need to ensure both automated enforcement and human review are deployed well. Triage, prioritization, and routing are key parts of any well-crafted system. To some extent, each service will probably need customization. Yes, each service has its own structure and focus that will guide the design of a functional system. But they are also not so unique that we need to reinvent the wheel for each one. There's an equilibrium somewhere between understanding the idiosyncrasies of each product or community, and drawing on shared concepts and approaches we can apply more broadly.”

QUEUEING AND WORKFLOW-MANAGEMENT TOOLS

Based on our interviews, most T&S teams start out on the most generalizable tooling solutions for workflow-management and queueing needs: some combination of providers like Zendesk and Jira, or similar SaaS (software as a service) solutions for ticketing and customer support. These tools are used to solve workflow needs, but more tailored solutions would help, particularly as platforms scale; a hacked-together workflow needs to integrate new tools or features to accommodate changing organizational structures, product needs, and review features.

This integration challenge was cited as a key tooling challenge for which open solutions could play a role. Open solutions for workflow management may look like a fit-for-purpose ticketing and enforcement interface, and may augment existing work to simplify prioritization and help organizations tackle queuing challenges for specific types of content.

When it comes to the interfaces used to review particular content and behavior, practitioners called out the importance of supporting reviewers' well-being. For instance, to mitigate the risk of exposing reviewers to grossly violent or abhorrent imagery, content-moderation systems may grayscale and blur images, allowing a reviewer to determine whether action can be taken without any further detail and context.

In addition, practitioners noted that regulatory requirements are increasingly directly relevant to case management. For instance, the Digital Services Act specifies how hosting services must provide users with a "statement of reasons" regarding removed content. To the extent these sorts of requirements drive some measure of standardization in T&S workflows, open solutions may also be helpful in providing off-the-shelf support for a variety of services.

MEASUREMENT

When violative content or conduct is detected and action is taken, that information is then fed back as a data point to further train automated detection mechanisms, build profiles of abusive behavior, and evaluate T&S' operations performance.

Services' data architecture may vary a great deal, and, thus, participants in our interviews suggested it would be difficult to generalize such a system. However, it could be possible to develop basic, open logging tools that build upon the enforcement infrastructure, labeling actions taken on various content.

Similarly, while different services track different trends, all need the capacity to produce at-a-glance analysis of their enforcement efforts. A general, open tool could be created to facilitate some of the most common elements that T&S teams might want to track across different types of content. Furthermore, such tools could incorporate common, best-practice metrics that support quality assurance; currently, public versions of these frameworks do not appear to exist, according to the practitioners to whom we spoke. Along with tracking metrics around violative content, practitioners also called out the opportunity for common frameworks for metrics around positive or "prosocial" uses of a tool (e.g., sharing of content meant to reduce intergroup hostility) and interaction with features designed to support such uses (e.g., features that remind people to be civil, and prompt them to consider the content their posting).

TRANSPARENCY

While much of their data will remain internal, services make data transparent to users, researchers, the public, and, increasingly, regulators. This transparency has associated tools needs.

Services build tools to report to a user that their content has been removed, and sometimes to provide appeals processes. They also might build dashboards that allow users to track content they have flagged, so that they can see whether it has been acted upon. These are highly customized to a service.

Services also build external transparency reports, reporting aggregate statistics and other information about their T&S processes to both the public and regulators. Aggregate transparency reports vary, too, in final presentation. However, services generally look to report common items like the number of pieces of content removed in a given category of content. Moreover, to the extent that legal requirements drive some amount of baseline convergence in reporting needs, a shared, open tool could provide a fit.

BUILDING A HUB FOR BEST PRACTICES

How might work across these areas be advanced? Practitioners did not have a unified view, but they pointed in a few directions. The overarching challenge is that there is no one-stop shop to consult to understand the tools that are available or to get advice on what considerations to keep in mind when implementing them. The creation of a single hub to drive a critical set of activities could be prudent.

A starting point for this work might be to simply compile and curate lists of the existing tools up and down the technology stack. Each tool might be described in a single place, detailing and benchmarking its capabilities, dependencies, and limitations, as well as what is required to implement it. In the absence of any organized effort to do this, some academic researchers have attempted to compile datasets, classifiers, and machine-learning (ML) tooling examples, but these efforts are nontrivial to identify or navigate.

We also heard from a number of T&S professionals that their teams and organizations would be interested in contributing open-source code for tools they have built in house. But many of these tools are not well suited to simply live on Github or another third-party site; rather, they require a responsible steward to maintain them over time, adapt them to specific deployments and communities, and keep them up to date with the current adversarial environment. For example, Discord wants to release an open-source version of its in-house rules engine at some point in the future, but hopes that a steward can be identified to maintain and support the tool to assist other organizations that choose to deploy it. Effective development for this and other open tools will require an organization staffed with technical experts, as well as T&S domain experts.

The opportunity to compile what already exists and support future contributions of open-source code speaks to the need for an institutionalized technical capability focused on T&S. Given optimal technical talent and organizational support, there is significant opportunity to build and contribute open-source and shared code and tools directly. This might take the form of releasing reference implementations of classifiers or other tools, or building and maintaining open-source tools, such as a workflow-management engine.

LOOKING OVER THE HORIZON

T&S is a dynamic field, and practitioners also encouraged thinking about how open tools might intersect with issues that are on, or just over, the horizon. We highlight three areas below.

ARTIFICIAL INTELLIGENCE

Over the last decade, the use of content classifiers built with machine learning and other artificial-intelligence tools dramatically changed the ability for services to identify new, potentially harmful content at scale. One of the most provocative comments we heard in our research was that we may be at another pivot point, as generative AI reshapes operational T&S over the coming years.

Researchers and services are already beginning to deploy generative AI systems to help at different parts of the T&S process.

- ▶ Microsoft researchers recently demonstrated how they used large-language models (LLM) to synthetically create a dataset of hate speech, which was then labeled and used to train a classifier.

⁷ Some vendors, such as Active Fence, have begun to provide templates for transparency reporting with these regulatory requirements in mind.

- ▶ Cohere.ai posits that LLMs will allow classifiers to be developed from much smaller datasets. Rather than requiring people to label myriad pieces of data to train classifiers, an LLM can use a small, labeled set to then label other data itself.
- ▶ OpenAI used GPT-4 to develop classifiers to identify harmful outputs of its system.
- ▶ Our interviews suggested that such systems could also be set up to take enforcement actions and provide an explanation to both the system operator and the user. The system could also be used to work in the reverse direction; that is, from a set of enforcement decisions, the system could be asked to distill what relevant rules apply to them.

We cannot evaluate how far and how fast these tools will impact T&S, but investment in open tooling for T&S should be attuned to these changes. Experts with whom we spoke emphasized that there will be a growing demand for tooling that allows people to use LLMs to rapidly spin up contextually appropriate T&S operations, including rules, systems, and classifiers. In other words, the AI may exist and be capable of significant impact on T&S operations, but tooling will be required to allow people to optimally use the AI for that purpose.

AI has also gained widespread attention for the role it plays in increased capabilities for anyone to generate synthetic content at scale. This growth of synthetic content has raised questions around so-called “deep-fakes” and how to detect them at scale. More generally, people can synthesize harmful content of all stripes. As noted above, datasets to train detection tools, reference implementations, and open classifiers could prove useful here.

EVOLVING SERVICE TYPES

T&S will need to evolve to adapt to the challenges of a number of new service types, including real-time communication, metaverse technologies, and the decentralized social web.

EXTENDED REALITY (XR) AND THE METAVERSE TECHNOLOGIES

XR technologies encompass both virtual-reality environments and augmented-reality tools that overlay digital objects on the physical world. These technologies have been most recently associated with the idea of a “metaverse” in which more of our Internet-enabled experience is done in real-time, embodied communication. On the one hand, real-time environments are nothing new to T&S. On the other hand, they have not been a predominant form of interaction. What’s more, the types of media objects involved (e.g., three-dimensional renderings) can require different detection tooling, and the behavioral and interactive elements may lead to different harms (e.g., simulated physical or sexual assault) that require different interventions. While the overall workflow of T&S may remain consistent, it might also respond to new signals based on the sensors embedded in the devices, raising new T&S and privacy issues.

One key way in which XR may demand a different focus for T&S tooling is the sheer variety of formats, spaces, modes of communication and interaction, group sizes, and environments that users may encounter. Community-moderation tools are already relied upon by services such as Discord, Reddit, Nextdoor, and Facebook, and these types of tools may become even more important in XR to accommodate the range of group dynamics that may unfold. XR may also heighten the need for user-specific T&S tools that would allow a user to filter certain conduct or content out of their experience. For example, Meta has incorporated a feature that garbles the voices of strangers on its Horizon Worlds service, so that talking to and engaging with strangers is something that a user can choose to avoid.

DECENTRALIZATION OF SERVICES AND USER CONTROLS

As we noted at the outset, T&S operations are not homogenous. While it is easy to think strictly in terms of centralized operations run by a service provider, there are services that are largely moderated by users themselves. What is more, in part due to a backlash against centralized control of online platforms, there are an increasing number of decentralized services evolving.

Mastodon is one example that illustrates some of the T&S challenges of the decentralized web. Unlike at Twitter or Facebook, there is no central Mastodon T&S team that moderates content across the service. In fact, it is impossible for such a team to exist, as each Mastodon server instance is responsible for moderating the content hosted on that server. At present, [the tools to do so are relatively rudimentary](#), and many of the data signals that T&S operations tend to use with centralized services (e.g., access patterns) are not necessarily usable on these services.

This is not simply an issue for Mastodon. For instance, the InterPlanetary File System, which facilitates distributed file storage, has worked to create an [optional hash list](#) for storage operators to facilitate blocking.

Participants in our interviews noted that the solutions for decentralized services may not simply be carbon copies of what works elsewhere. In particular, to the extent that the operators of individual nodes in the network are hobbyists, they may need very different sorts of tooling.

Relatedly, more decentralized architectures may open up opportunities for new sorts of user controls created by a range of parties. For instance, the Bluesky social media service is being designed so users can use different clients; in addition, developers can create different [algorithmic feeds](#) and users can select from them. It also allows people to create moderation labels that developers and users can use to support more customized content-moderation choices. To be clear, user controls do not necessarily depend on decentralized architectures for services; for instance, the [Block Party](#) is a third-party tool for Twitter users, helping them more easily filter out possible harassment. The point here is simply that new, more decentralized service architectures may expand these opportunities.

AGE ASSURANCE

Services may engage in verification processes for users for a variety of reasons, and age assurance is a subset of user-verification issues, referring to methods services can use to estimate a user's age with varying levels of certainty. While legal requirements around age assurance remain controversial, they are increasingly common, and service providers are having to reckon with how they will adapt. Determining a user's age with more granularity may trigger requirements and opportunities to apply certain safety measures; at the same time, collecting data necessary to make that determination raises a host of privacy and security questions.

Several experts in our interviews pointed to the potential benefit of open solutions in this area. We heard three key areas in which open solutions could be helpful.

Interoperability: Service providers and users could benefit from an organizational or technological solution that creates interoperability, so that a user who has their age assured via one service does not need to repeat the process at other sites. [Relevant efforts are under way in Europe to create a shared solution](#) for age verification that effectively allows users to reuse an age check across participating providers in a privacy-preserving way.

Age-inference models: One age-assurance method is through inference based on data a service already has. Just as with content classifiers, one can imagine tools that estimate age based on a common set of inputs. For instance, researchers created an open-source tool to infer demographics based on public Twitter data,

and one could imagine a broader classifier for public text on social media that estimates age. Along with raising similar challenges as classifiers generally, the challenge here is that many attempts at estimation will rely on private data, and solutions that work for one provider's data architecture may not be easily transferable.

Age inference can also be done via tools that analyze a person's face. Open-source tools to do this already exist. While robust open solutions could help lower the cost of this method, it would not address the fundamental privacy challenge involved in collecting that biometric information.

Nonprofit clearinghouse: A much more ambitious approach would create an entity that would be entrusted with collecting relevant age-assurance information from a user. This would not eliminate the privacy and security concerns with collecting user information for age assurance. However, it would attempt to address them through an intermediary that is noncommercial and entrusted to the public good, separate from the government or any companies. It could necessarily require significant technical expertise to get this right—but, even more so than with some of the concepts above, the institutional challenge here is the tip of the spear.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This report was authored by Proteus Strategies, a boutique policy consulting firm based in San Francisco, California, founded by Derek Slater and Betsy Masiello. Collectively, the Proteus team has more than thirty years of experience managing communications and public policy efforts on issues such as online safety, data privacy and security, and open infrastructure from inside leading technology companies. This work was supported by the William and Flora Hewlett Foundation. The authors are grateful to the members of the Atlantic Council Task Force for a Trustworthy Future Web for providing their feedback and perspectives.

APPENDIX**OPEN AND SHARED T&S TOOLS REFERENCES**

The following is a compilation of open or shared tools referenced above. As noted, there are a wide variety of open tools available, so this is not comprehensive.

HASH-MATCHING CAPABILITIES	<ul style="list-style-type: none"> ▶ Facebook’s PDQ, TMK+PDWF, and Hasher-Matcher-Actioner (GitHub repo) . ▶ Other referenced perceptual hashing tools like pHash are widely available, implemented in different forms—see, for example, Github repos. ▶ For CSAM, matching tools include <ul style="list-style-type: none"> — Microsoft PhotoDNA; — Google CSAI Match; and — Cloudflare CSAM scanning tool.
HASH DATABASES	<ul style="list-style-type: none"> ▶ CSAM databases are managed by groups like NCMEC, Internet Watch Foundation, Thorn (its Safer Database), Interpol, the Dutch Expertise Bureau Online Kindermisbruik (EOKM), and the Canadian Centre for Child Protection (Project Arachnid). ▶ GIFCT operates a hash database of violent extremist content. ▶ StopNCII.org’s database pertains to nonconsensual intimate imagery (aka “revenge porn”).
CLASSIFIERS	<ul style="list-style-type: none"> ▶ Google Content Safety API—focused on child-abuse material. ▶ Bumble’s Private Detector. ▶ Jigsaw PerspectiveAPI and ConversationAI-Moderator tool. ▶ Unitary Detoxify. ▶ Yahoo Open NSFW. ▶ GitHub features many relevant libraries—see, for example, pages for hate-speech detection tools and violence detection. ▶ GitHub and Hugging Face similarly have many relevant models and datasets—see, for example, a search for “toxic” to find relevant detection tools on Hugging Face. ▶ OpenAI Moderations endpoint.
REFERENCE IMPLEMENTATION AND GUIDES FOR ML/AI DEVELOPING TOOLS	<ul style="list-style-type: none"> ▶ As one example, TensorFlow published its guide to use for content moderation, as one among many platforms. ▶ Various GitHub libraries provide overall guidance—see, for example, “content-moderation-deep-learning.”
OPEN DATASETS	<ul style="list-style-type: none"> ▶ Both Github and Hugging Face contain relevant datasets for content detection—see, for example, datasets related to toxic content. ▶ Google contributed a deepfake dataset. ▶ Ubisoft and Riot announced a collaboration to build a shared dataset.
ENFORCEMENT	<p>Rules engines, queueing, and workflow-management tools are available in various open forms, but we did not find items tailored for T&S.</p> <ul style="list-style-type: none"> ▶ GitHub has a number of rules engines, but none tailored to T&S. ▶ Open-source tools like osTicket have general-purpose ticketing and case-management capabilities, much like Zendesk, Jira, and similar tools.

ANNEX 3

SCALING TRUST ON THE WEB

RESPECTING CHILDREN AS RIGHTS HOLDERS

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

ANNEX 3

RESPECTING CHILDREN AS RIGHTS HOLDERS

TABLE OF CONTENTS

Children’s Safety	3
Children’s Rights	4
Tension Between Children’s Safety and Other Rights	5
Approaches to Youth Engagement	6
Key Concepts in Online Children’s Rights and Safety Policy	7
Artificial Intelligence (AI)	7
Child Sexual Exploitation and Abuse	8
Duty of Care	9
End-to-End Encryption	9
Ephemerality	9
Mental Health	10
Parental Controls	10
Transparency Reporting	11
Conclusion	11
Authorship & Acknowledgments	11

A TRUSTWORTHY FUTURE WEB MUST REFLECT THE RIGHTS—AND PERSPECTIVES—OF CHILDREN.

→ Digital technologies, the entities that provide them, and the laws and policies that apply to them are all generally the work of adults. Yet, adults and children alike use (and are otherwise impacted by) those technologies, must operate within those policy frameworks, and must have their views and respective online practices reflected in them. A growing body of research shows that involving children directly in decisions that impact them (in an age-appropriate way) is key to identifying their best interests, and to effectively understanding and addressing children’s diverse needs.¹ However, children have traditionally had little direct say in the development of the technology policies and digital products that affect them.² Instead, their interests—if they are considered at all—are largely represented by adults (indeed, even in the drafting of this piece).

Policies and products aimed at protecting children’s safety have increasingly played a pivotal role in influencing trust and safety practices within companies, as well as driving forward new laws and regulations (some proposed, some adopted) based in protecting children’s safety.³ Combined with the absence of actual children from policy conversations, this singular focus on children’s safety has demonstrated a challenge inherent to establishing safety as the foundational premise of tech policy development: it can lead to the violation of other crucial rights that children enjoy, do so without creating any space for their input or involvement in making that tradeoff, lead to counterproductive outcomes, and result in violating the rights of whole other communities of rightsholders.

¹ Emily Weinstein and Carrie James, *Behind Their Screens: What Teens Are Facing (and Adults Are Missing)* (Cambridge, MA: MIT Press, 2022).

² Compare tech policy to, say, education for women and girls, or the climate crisis. By contrast, those are policy issues on which young people, who are directly (and negatively) affected by adults’ policy choices, have insisted on making their voices heard, turning activists such as Malala Yousafzai and Greta Thunberg into household names while they were still teenagers.

³ For a brief history and layout of the online-safety ecosystem, see: Anne Collier, “The Child Online Safety Ecosystem: A Look at the History, Education, Content Moderation and Developments around the World” in Kalinda Raina, ed., *Children’s Privacy and Safety* (Portsmouth, NH: International Association of Privacy Professionals, 2022).

In our collective time working on these issues, we have noticed four common themes in tech policy discussions involving children: the assumption that digital technologies are (primarily or solely) detrimental to children; that children can be considered as a homogeneous group; that children’s safety from physical, mental, and (especially) sexual harms is achieved through limiting children’s access to digital media and devices, rather than through empowering harm-reduction approaches or prioritizing the full range of digital rights children hold; and a preference to address online child-safety concerns through criminal law enforcement and increased digital surveillance (by parents, companies, and state agencies such as schools and police).⁴

In this annex, we aim to expand the aperture, giving space to considerations that can otherwise be excluded when the narrative is restricted to child safety, and hopefully illuminating underexplored areas for future attention and investment. We seek to reframe the discussion by centering children’s rights and agency, and treating safety as part of a constellation of rights that children enjoy, and that must coexist with the other rights of children as well as the rights of adults. We begin by defining what we mean by these two terms (“children’s safety” and “children’s rights”) and noting the tension between them in existing policy discussions. We then explore lessons that have been learned in both tech and non-tech spaces regarding methods for supporting the participatory inclusion of children safely and effectively. We close with a wide and illustrative range of policy areas in which children’s rights can, and should, be considered, and where their inclusion could be operationalized.

We would like to acknowledge from the outset that the regulations, legislation, and citations grounding this article reflect expertise in the evolution of these questions within the European Union (EU), the United Kingdom (UK), and the United States. Child safety and children’s rights are key issues to be examined and illuminated around the globe. We would like to note the importance of giving greater attention to the evolution of these concepts across other regions, and encourage investments in supporting a more equitable body of research in order to better inform the wide range of stakeholders making policy and product determinations with regards to children’s safety and rights.

CHILDREN’S SAFETY

“Child safety” is an umbrella term that can mean different things to different people (children included). The phrase might primarily evoke defense against child sex-abuse material and exploitation (CSEA, a term that includes imagery depicting child sex abuse, child sex trafficking, and sextortion, among other concepts). However, its meaning goes well beyond sex-related harms and can extend to safety from physical violence, threats, hate speech, and harassment (e.g., parental abuse, state violence, cybersecurity-related harms) or psycho-social well-being (e.g., “screen addiction,” cyberbullying, eating disorders, self-harm), among other concerns.

“Child safety” is not only a broad term, but also a politically charged, culturally nuanced concept. Because “children’s lives have become digital by default,” “online safety” for children spurs fierce debate and passionate focus among adults seeking to establish broad and enforceable rules for a constantly shifting, highly personal paradigm. Tradeoffs regarding agency, autonomy, governance, privacy, security, and a range of related values will never be simple or universal. However, the process of identifying and debating tradeoffs (as well as looking for creative ways to avoid tradeoffs and optimize for multiple values) can be grounded in the practices that have been developed over decades to grapple with exactly these complexities and provide a coherent foundation for negotiation and consensus building. That foundation includes long-standing frameworks of human-rights law.

⁴ This preference embeds a further assumption: that parents and the state do not harm children’s safety or infringe upon their rights.

CHILDREN'S RIGHTS

Children are humans, so children share in the human rights enjoyed by adults, in addition to specific rights based on their status as children. Human rights apply online as well as offline.

To define children's rights, this annex refers to the United Nations (UN) Convention on the Rights of the Child (CRC). The CRC has been ratified by every UN member state except the United States.⁵ That means those 196 countries have voluntarily agreed to abide by the CRC, an obligation that extends to these countries' laws and policies regarding digital technologies. In 2021, the UN published guidance, called General Comment 25, to states regarding their CRC obligations in the digital environment. That guidance has had differing impacts in different stakeholder settings, with the greatest traction occurring in European, Australian, and UK governmental debates. Its impact in the United States is limited, in part, because the (typically private-sector) entities that actually create the digital environment are more focused on the binding laws applicable to them than on nonbinding guidance directed to states, though European legislation such as the General Data Protection Regulation (GDPR) and Digital Services Act (DSA) will have growing impact even on US-based companies' child-protection practices.

The CRC recognizes that children have the right to protection from violence, abuse, and exploitation (Articles 19, 34, 35, and the Optional Protocol on the Sale of Children, Child Prostitution, and Child Pornography (OPSC)). Children's rights go well beyond safety, however. Scholars have categorized their rights under what they've characterized as the "three Ps": protection, provision, and participation rights, which, under General Comment 25, have their digital counterparts, including the right to express their views on matters that affect them (in accordance with the child's age and maturity) (Article 12); rights to privacy (including in their correspondence) (Article 16) and freedom of expression (Article 13); the right to seek and receive information (including access to mass media) and the right to education (Articles 13, 17, and 28); freedom from discrimination (Article 2); rights in the justice system (Article 40); and freedom from all forms of violence (physical, mental, sexual), exploitation, and trafficking (Articles 6, 19, 24, 32, 34, 35, 36, and 39).

All of these rights are universal, non-hierarchical, indivisible and interdependent, equal and nondiscriminatory. One set of rights cannot be enjoyed without another, which makes them all equally important—and makes the three categories of protection, provision, and participation a helpful framework for upholding them in balance with one another. Implementing decisions based in human-rights principles is, and remains, a challenging task, but new guidance is emerging and is worth noting. For example, the UN Human Rights B-Tech project provides guidance for technology companies about how to carry out human-rights due diligence, including human-rights impact assessments. UNICEF has worked with UN Human Rights to produce a special briefing on how children's rights can be considered by technology companies as part of the human-rights due-diligence process, which will be published later this year.

TENSION BETWEEN CHILDREN'S SAFETY AND OTHER RIGHTS

Children's rights are co-equal, meaning that, in theory (as General Comment 25 notes), protecting their right to be safe online must not come at the expense of their other rights, for example, safety at the expense of privacy. This is a difficult balancing act in any context, and especially within the context of ever-evolving digital spaces in which the threats and opportunities for children can shift and scale with astonishing speed. While most agree that preventing harms against children in their digital environment is imperative, there is a lack

⁵ While the United States has not ratified the CRC, it has ratified the CRC's Optional Protocol on the Sale of Children, Child Prostitution, and Child Pornography. According to Human Rights Watch, children's rights are poorly protected in the United States, prompting scholars to posit an alternative model of parental rights within US law that promotes a broader range of children's present and future interests.

of consensus among stakeholders as to how to prevent harms while protecting children’s agency and rights of participation, and often a lack of technological solutions that could immediately address harms even if a policy determination were made. The interpretation of long-standing children’s rights also constantly evolves as a component of broader digital transformations—for example, traditional prohibitions against child labor are now being considered within the context of a raging “creator economy.” This creates an ongoing tension in tech policymaking, trust and safety tooling, and broader digital-product development.

Children are frequently positioned within policy debates as dependents or subordinates of those who hold rights over them (such as parents or the state). Deep tensions exist around the question of where to set the boundaries of parents’ and children’s own respective agency and accountability. Moreover, children have historically emerged as the framing (one that parents themselves do not necessarily condone) for much larger normative debates within communities about highly sensitive political topics and evolving cultural norms around sexuality, gender, and religious or ethnic identity. Finally, proponents of approaches that center children’s safety as one of many competing rights that must be considered can face a high political cost, as well as personal attack.

Alongside recent debates about cyberbullying, Internet addiction, and data privacy, one of the most significant and recurring threads in tech policy debates involving children centers on how to keep children safe from CSAM (especially CSAI) and impede the vast scaling and potential normalization of such material.⁶ This policy priority has inspired legislation and regulatory proposals in the United States, EU, China, and elsewhere, although legal frameworks vary in robustness in terms of definitions and effective criminalization of the perpetrators of CSAM crimes. While most stakeholders agree that preventing harms against children in their digital environment is imperative, they lack consensus about how to do so while protecting children’s agency. This disagreement not only creates an ongoing tension in tech policymaking, but also reflects a more fundamental normative challenge. Stakeholders do not necessarily operate from a baseline agreement that children are rights holders in terms of their own rights and agency, nor do they share a baseline agreement on the extent to which adults’ rights can, and should, be limited or implicated.

Great opportunity arises when this tension forces a recognition that competing rights can, and must, be balanced. For example, in August 2021, Apple announced a plan for new iPhone child-safety features that drew controversy due to their negative impacts on privacy and other rights of both adult and child iPhone users. An executive director at the National Center for Missing and Exploited Children (NCMEC) privately congratulated Apple for “prioritizing child protection” and described those who expressed concerns as “the screeching voices of the minority.” This is an example of the political challenges that rights advocates can face in debates around child safety.

After public pushback, Apple dropped the most controversial (and privacy-intrusive) component of its plan, and revised another feature to give child users more privacy and agency, responding to concerns that the original design would jeopardize the safety of child users with abusive parents. While Apple’s original plan subordinated some rights (of children and adults) to children’s right to be free from CSAE (especially CSAI), its revised plan reflected a more non-hierarchical approach to its users’ rights and an understanding of child protection that includes agency, digital privacy, and safety from family abuse, not just safety from CSAE. (That said, the revised plan was not well received by NCMEC.)

Notably, Apple’s review benefited from several key elements: existing normative agreement and legal standards regarding the unacceptability of the production and dissemination of CSAI; knowledge exchanges that reflected deep expertise on critical policy components, as well as complex technological questions; and in-

⁶ For a deeper analysis of this topic, see *Annex 4: Deconstructing The Gaming Ecosystem*.

creased opportunity for input by a more extensive group of advocates, companies, and regulators than Apple had originally consulted. However, the perspective of one key set of stakeholders remained notably absent from this collaboration: children.

APPROACHES TO YOUTH ENGAGEMENT

In some countries, youths are actively engaged, to some extent, on relevant issues such as privacy, consent, and data sharing in the digital realm, although ample opportunities remain to expand such engagement throughout the world. Many children are adaptable, nimble, tech-fluent digital natives, often with well-formed opinions about the digital environment's impact on their rights. Their perspectives and solutions-based thinking can vary widely from those of policymakers who may overwhelmingly represent a significant generational divide from them. Youths' voices not only deserve to be elevated within debates around their rights and safety in the digital realm, but their lived experience with social media is needed for adult intelligence gathering and policymaking; they may also add significantly to the richness and creativity brought to bear in identifying strong paths forward.⁷ This is true for decision-making spaces across the ecosystem, be it in informing governmental debate, within corporate decision-making processes, or setting research and advocacy priorities. The UNICEF and Lego RITEC project, which sets out a framework for integrating the well-being of children into digital design, centers even young children's own participation in defining what it takes to maximize children's well-being.

In considering approaches to youth engagement on the subject of children's rights and safety online, it is imperative to recognize the diverse experiences and perspectives of young people, which can be shaped by innumerable factors and may include geography, race/ethnicity, religion/faith, socioeconomic status, sexuality, gender identity, cognitive development, and (significantly) levels of trauma, among many others. While there is no one-size-fits-all approach, there are existing methodologies and practices for operationalizing youth engagement effectively and ethically in tech design and policymaking (both in and outside of the technology sector), as well as innovative new public-, private-, and cross-sector initiatives.

One established consideration in youth engagement is to examine the intended goals and outcomes of youth-engagement efforts. Should such efforts be activated to raise youth awareness of online harms, which might include different learning and teaching strategies, such as social and emotional learning? Should such efforts prioritize youth engagement, which might include youth development, collective empowerment, and/or systems changes? Different goals necessitate different approaches.

When engaging with children on topics involving online harm, it is also important to accommodate children's own history of, or ongoing experiences with, violence. This should be based in an understanding of the typology of different harms (e.g., armed conflict, familial, racial, or sexual violence, etc.), children's possible experiences of multiple overlapping harms, and these harms' potential impacts on the child. For example, how might a youth-engagement strategy approach a child or teen who has been sexually abused, and who, as a result of this trauma, acts out sexually inappropriately toward other youth online? Trauma-informed frameworks can provide an additional lens through which such learning or engagement strategies can be developed.

Depending upon which strategy is most effective for determined outcomes, it is valuable to include educators, mental-health professionals, and child-development specialists alongside youth (or to review resources in these areas and request consultations, if it is not possible to include youth-serving professionals during actual discussions). Such inclusion will help ensure any discussions are appropriately scaffolded for youth (using the abovementioned strategies and frameworks) depending upon youths' cognitive development and existing experiences with trauma, among other factors.

⁷ See, supra note 2

KEY CONCEPTS IN ONLINE CHILDREN'S RIGHTS AND SAFETY POLICY

The following concepts are key areas of current inquiry in tech policy and product-development debates that hold particular salience for children's rights. Each raises underexplored avenues for future research, presenting an opportunity for ongoing, innovative investment framed through a rights-centric lens. For each concept, a full discussion of all of the rights and tradeoffs involved is beyond the scope of this summary-level document. However, the CRC-enumerated rights listed above can serve as a jumping-off point, particularly the four principles grounding the CRC: non-discrimination; a child's best interests as a primary consideration; survival and development; and the right to be heard.⁸

ARTIFICIAL INTELLIGENCE (AI)

AI holds great promise for helping children stay safe and exercise their rights. AI can be used to locate information efficiently, and for generating (or translating) text, sound, video, and images. Thus, children can use AI to learn, express themselves, create art, play, and socialize with others. AI-based tools are already crucial to the automated detection and review of potentially abusive content online at massive scale (although algorithms can replicate bias, and the use of CSAI to train AI tools is controversial and legally dubious). Offline, AI is starting to be used in cancer detection and drug discovery, as well as to aid the work of professions that help protect children, such as legal services and refugee services.

At the same time, AI's perils are already well documented. Children's photos or other personal information may be ingested into (or output from) AI systems. Children may be exploited or bullied by using their images in deepfakes. AI-generated misinformation and disinformation may mislead and misinform kids; synthetic media might fool them into falling prey to scams or hacks. Depending on how they are used, generative AI tools can support and/or undermine a child's education. In the juvenile-justice system, decisions may be made using algorithmic tools that replicate systemic biases. Given these risks, claims about AI require exceptional vigilance, especially claims of "AI that will help children." Stakeholders should approach AI project proposals with skepticism, and must demand satisfactory answers about ethics, privacy, safety, and risk mitigation.

CHILD SEXUAL EXPLOITATION AND ABUSE

Perhaps uniquely as types of content go, depictions of child sexual abuse are illegal basically everywhere, making CSAI a common compliance concern worldwide. Online CSAI is an enduring problem, despite decades of concerted technological and political interventions. This is tragic and yet unsurprising, since the instrumentalities for CSAE offenses are the core functions of the open Internet: transmitting, storing, and accessing files, and communicating with other users.

Online CSAI sharing and solicitation is a deeply complicated and multifaceted issue. For starters, the problem eludes precise quantification, with debates over the reliability of official numbers of CSAI reports by online services and the reasons why they have ballooned lately. Plus, it is not only pedophiles who share CSAI. The sharer's motivation may be humor or outrage, or the child depicted may have produced and shared the image initially. Imagery created by the child depicted, called self-generated CSAI (SG-CSAI), raises a range of complex issues. Some SG-CSAI involves a teenager who is of age and consenting. Some is harmful, as with grooming, sextortion (someone extorting the child into creating and sending CSAI), or the nonconsensual

⁸ For a deeper analysis of this topic, see *Annex 1: Current State of Trust and Safety*; *Annex 2: Building Open Trust and Safety Tools*; and *Annex 4: Deconstructing The Gaming Ecosystem*.

sharing of an image that was originally created and sent consensually (nonconsensual intimate imagery, or NCII, also called “revenge porn”). SG-CSAI is still illegal, so consenting teenagers may risk prosecution for documenting legal sexual activity.

Service providers at multiple levels of the tech stack have devoted significant resources to detecting CSAI. Many major tech companies scan their users’ files with so-called “hash” technologies (such as Microsoft’s PhotoDNA). New, unknown images and livestreamed abuse pose greater technical challenges for automated detection than do known images. Similarly, exploitation offenses (e.g., grooming, enticement, and solicitation) pose a different technical challenge than the detection of known CSAI. While text-based classifiers for CSEA have been developed, finding matches for known imagery is more straightforward than determining whether a text or voice interaction is abusive. Different CSAI and CSEA scanning technologies vary as to their accuracy and the implications for user privacy.

Through automated scanning for known CSAI, major online services detect large volumes of CSAI every year. While this technology aids children, it has a significant privacy impact. Many tech companies scan voluntarily, but policies to compel scanning for CSAI have been proposed in jurisdictions such as the EU and India. These proposals can conflict with constitutional (e.g., Fourth Amendment) or statutory (e.g., General Data Protection Regulation, European Convention on Human Rights) privacy protections; even voluntary scanning may run afoul of some privacy laws absent a carveout. Likewise, privacy laws may impede platforms from sharing information with each other about users who use multiple services to share CSAI or exploit children.

CSAI scanning helps to detect and stymie horrific sex abuse of children. However, scanning all online content and communications is a significant privacy intrusion that also has effects on other rights such as free expression and association. Scanning mandates for CSAI also pose a threat to human rights (including children’s) because authoritarian governments could also require online services to scan for other types of content: political dissent, religious expression, LGBTQ+ (lesbian, gay, bisexual, transgender and queer) content, etc. Because automated scanning for CSAI does not work in end-to-end encrypted environments, which are relied upon for safety and privacy around the world, legislative proposals that overtly or effectively require scanning for CSAI pose a risk to online service providers’ ability to legally offer end-to-end encryption.

Policy proposals to combat CSAI and child sexual exploitation often reflect an attitude of “tech solutionism”: expecting technical interventions to fix the larger, sticky societal problem of CSAE. Anti-CSAE policy proposals tend to focus on law enforcement and criminal punishment. Comparatively few have focused on prevention and support—for example, by investing in child-protection systems, sex-education and CSAE-awareness programs, and housing stability for at-risk youth. A comprehensive response requires a victim-centered criminal-justice system, coupled with prevention measures and support for victims. Victim-focused policies are more rights respecting than those that center the interests of criminal law-enforcement agencies. However, policymakers may prefer approaches that rely on self-funded tech-company initiatives, rather than government support. Prevention- and support-centric policy models for mitigating CSAE are, thus, an area in which philanthropic investment could expect to meet with a favorable response from both governments and tech companies.

DUTY OF CARE

The concept of the duty of care is a common-law term that comes from the law of negligence and is used to impose a standard of care on technology companies to avoid careless acts that could foreseeably harm children. Duty-of-care obligations, some specific to minors, are a key part of the EU’s new Digital Services Act and the UK’s pending Online Safety Bill. The services covered by such duties are typically those accessible to, or likely to be accessed by, children. Definitions of the duty of care vary, but typically require only collecting data on young users that are required for a service to function, setting strong default privacy settings, and avoiding deceptive or addictive design features.

In practice, the duty of care effectively requires age assurance to determine which users are children (or, alternatively, verifying all users' ages to ensure compliance, potentially imperiling most users' privacy), meaning it implicates the same rights as age assurance (see [Annex 2: Building Open Trust and Safety Tools](#)). Duty-of-care bills have also been criticized for letting regulators and tech companies decide what is in children's best interests, which jeopardizes children's access to online content that might, in fact, be crucial for actual children's well-being (mental or sexual health resources, LGBTQ+ resources, etc.). It is unclear how duty-of-care obligations will apply to end-to-end encryption (E2EE). Fear of liability could dissuade providers from offering E2EE services to children, or at all; conversely, duty-of-care policies that require the most privacy-protective settings by default for children's accounts could be read to mandate default E2EE for child users.

END-TO-END ENCRYPTION

End-to-end encryption is a technology for protecting data privacy and security by encoding data so that they can only be decoded by the sender and intended recipient(s). A recent report by Child Rights International Network (CRIN) and Defend Digital Me (DDM) analyzes the complicated interactions between E2EE and children's various rights—some positive, some negative. As the report describes, E2EE messaging enables children to exercise their rights (free expression, privacy, etc.), and protects children's lives and safety by keeping outsiders (such as an oppressive government or abusive parents) from monitoring their communications.

However, E2EE can also be used to violate children's rights, because it complicates the detection of abusive interactions with, or involving, children (although research shows that providers have other means of detecting abuse in E2EE settings). E2EE's usage in CSAE offenses, in particular, has prompted regulators and other stakeholders to call for regulating E2EE to enable investigatory access to users' communications. These proposals implicate users' rights (including children's rights), and would undermine E2EE systems' privacy, security properties, and user expectations. The question of how to mitigate the use of E2EE services for CSAE, without detriment to children's and all users' rights or harming digital security, is an ongoing area of intense and emotionally charged debate.

EPHEMERALITY

Ephemerality is a functionality of multiple popular messaging and social media services commonly used by children, such as Snapchat, Instagram, and WhatsApp. Ephemerality can be fundamental to how an app works, or it can be an optional function that users can choose to turn on. Ephemeral messages, photos, videos, or collages (e.g., Instagram Stories, WhatsApp Disappearing Messages, Snapchat Snaps and Chats) disappear after a set amount of time. Whether that period is triggered upon send or upon receipt/viewing depends on the app. The time length may be set by the app, or the app may give the user a range of time periods to choose from. Generally, ephemeral content disappears for everyone including the sender, but the sender may have the option to save the content.

Ephemeral functionalities enable children to express themselves spontaneously, and to communicate privately and securely, with less worry that what they share will stay online, spread, or linger in someone else's message history. However, ephemerality can enable abuse: it is harder to determine after the fact whether an account posted harmful content, and there is a limited time period in which a user can report content before it disappears. Plus, ephemerality is not bulletproof: recipients may screenshot or otherwise preserve ephemeral content (although some apps try to defeat screenshotting or notify the sender about it). This may pose a risk to children who share sensitive information without realizing it could still be preserved and shared (e.g., with a parent or school).

MENTAL HEALTH

There is significant debate and [public concern](#) about social media's effect on children's mental health and brain development. Research findings to date do not paint a consistent picture. Some research has shown significant differences by [geographic region](#) in social media's relationship with youth well-being, and (for reasons that are not yet understood) social media's impacts on individual youths' mental health can be [highly heterogeneous](#). In the United States, studies indicate that teenagers' [mental-health crisis](#) is not [clearly caused](#) by social media; indeed, some [research](#) has shown positive effects. Other research has examined behavioral interventions' impact on teen mental health, such as limiting screentime, also with [mixed results](#).

There is ample opportunity for additional research, with regard to both geographic distribution ([many studies](#) to date are from the United States and Europe) and age (outcomes for teens should not be extrapolated to younger children). Additional research is important because, despite the complex picture that existing research has painted, policy proposals [commonly assume](#) that social media is an unalloyed harm to children. It also suggests that technological or [product fixes](#) can solve problems that originate entirely outside of a platform—for example, within a school. That assumption leads to [laws that](#) impact children's rights online, such as by imposing age-related restrictions on children's access to social media (with largely arbitrary age thresholds). Social media's role in children's mental health is an area in which philanthropic investment in research, especially in the Global Majority, could make a significant impact.

PARENTAL CONTROLS

Parental controls are intended to give parents some say in shaping the digital environment their child experiences. Controls may apply to the form and/or the substance of a child's Internet use: both what information they access online (content restrictions) and how they access it (what times of day, for how long, how frequently). Over time, parental controls have evolved from something that policymakers incentivized (in the early years of the Internet) to something that policymakers are [increasingly attempting](#) to mandate for online services in jurisdictions such as the United States, UK, and EU. Requirements vary widely, but often include a dashboard that provides parents or guardians easy access to set default privacy and security settings, set screentime limits, and/or filter out certain content for their children. In some cases, [laws](#) require platforms to give parents or guardians access to their children's account to view their activity and communications, or to track their location.

The issue of parental controls and the emphasis on parent interests illustrate how policy discussions can be misaligned—if not at odds—with children's interests. Ideally, parental controls help supportive parents keep their children safe online and guide them into healthy habits. However, from a children's-rights perspective, parental controls may also [catalyze certain harms](#), depending on a given child's situation (as noted in the Apple example above). They [may exacerbate](#) harm to vulnerable children, such as those in abusive home environments or who would be at risk if a parent learned of their sexuality, gender identity, religious views, etc. More broadly, the effectiveness of parental controls—and their effects on children's development, cognition, and other interests—are not yet well understood. Nor is the impact that parental access to a child's online activity has on children's understanding and exercise of their privacy, free expression, and other rights. For example, how might the proliferation of parental controls shift an expectation of privacy among generations of youth? What impact does a reduced sense of privacy have on the child's cognitive and social development, civic engagement, and collective empowerment? As policy mandates multiply, parental controls will increasingly present a fertile opportunity for philanthropic investment in research.

TRANSPARENCY REPORTING

Lately, regulators often propose making major online services periodically report to the government about their internal policies, practices, and design features for keeping users safe on their services. [Some](#) proposals are child specific, whereas [others](#) apply to all users. Some, like [the EU's new DSA law](#), create programs to fund studies on technology's effects on children, and/or require covered services to grant researchers access to their internal data for independent study.

Transparency policies could greatly improve public understanding of popular online services' effects on child users. Enabling third-party research could reveal important insights, inform best practices, drive evidence-based policymaking, or improve online services' [often-disappointing policy](#) enforcement against abuse. However, there are risks: inadequate safeguards for researcher access could affect children's privacy. Requiring detailed public reports on services' internal practices could [give abusive users a roadmap](#) to circumvent them. These mandates also risk [creating a vector](#) for state censorship if services feel pressured to change their content-moderation and child-safety programs to better suit the state's preferences (a particular risk for content such as LGBTQ+ content or sexual-health information that some states consider harmful to children).

CONCLUSION

Ensuring children's safety and upholding children's rights require a holistic approach that encompasses the full range of children's rights, including participation, provision, and protection. Policymakers, researchers, advocates, and industry leaders must not only involve young people in decision-making, but also consider children's lived experiences within digital transformations, now and in the future. It is imperative that a wider range of stakeholders be supported to build models that can lower the barrier to incorporating rights-respecting frameworks and inclusionary models in the development of policies and products that impact children. This is particularly critical with regard to Global Majority communities, where existing inequities within policy, product development, and research risk being doubly visited upon children.

Emphasizing the consideration of the entirety of children's rights does not diminish in any way the critical importance of protecting them from harm within a quickly evolving threat landscape. It does, however, improve the likelihood that future iterations of online spaces will provide space and opportunity for children to benefit from online spaces where they can not only be safe, but also be empowered to learn, explore, play, grow, and evolve.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This annex reflects contributions from the following members of the Task Force for a Trustworthy Future Web: Lauren Buitta, Girl Security; Emma Day, DFRLab; Riana Pfefferkorn, Stanford Internet Observatory; and Leah Plunkett, Harvard Law School. It also reflects contributions from contributing experts: Sara Grimes, University of Toronto, and Anne Collier, the Net Safety Collaborative. This report does not represent the individual opinion of any contributor, member of the task force, or contributing organization. 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, as well as to Lauren Quittman of Duco Experts. Finally, the contributors would like to thank John Perrino, policy analyst at the Stanford Internet Observatory, for providing a briefing note on common policy approaches to children's online-safety issues and for allowing the task force to incorporate portions of the briefing note into this annex.

ANNEX 4

SCALING TRUST ON THE WEB


DECONSTRUCTING THE GAMING ECOSYSTEM

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

ANNEX 4**DECONSTRUCTING THE GAMING ECOSYSTEM****TABLE OF CONTENTS**

Introduction	2
A Categorization of Companies in the Gaming Ecosystem	3
Hardware	3
Content	4
Purchased and Downloaded Games	4
User-Generated Content	4
Content Creators	5
Game Developers	5
Development Technology	6
Publishers	8
Storefronts	8
Adjacent Communities and Platforms	10
Current Gaming Business Models	11
Market Leaders in Gaming and Their Countries of Origin	12
Global Investments in Immersive Technologies and Gaming Platforms	12
Implications for Further Study and Exploration	13
Trust and Safety	14
Learning from Gaming’s Challenges and Unique Perspective	14
Learning from Gaming’s Innovations and Approaches	14
Illuminating Existing Gaming Content Moderation and Trust and Safety Layers	15
Regulation and Governance of the Gaming Sector	16
The Rise and Impact of Generative AI in Gaming	17
Finding Signal Amid New Technology Hype Cycles	17
Geopolitical Implications and Influence	17
Conclusion	18
Authorship and Acknowledgements	19
APPENDIX Gaming Ecosystem Matrix	20

INTRODUCTION



Ongoing global debates over the internet tend to focus on major social media companies like Meta, Google, or Twitter. This inevitably shapes discussion and ideation around approaches to content moderation, trust and safety, and even future technology. In the past year, talk of the “metaverse,” the rise of distributed technologies, and a sudden explosion of consumer-accessible artificial-intelligence (AI) applications have begun to broaden conceptions of what our digital world might resemble. While there is still a great deal of uncertainty around exactly which of these trends is most likely to dominate future digital spaces and how, the companies building and experimenting with everything from extended reality (XR) to AI are in many cases not new, but, rather, are converging from adjacent industries.

One such industry is gaming—which has long served as a significant piece of the growing digital ecosystem but, for a number of reasons, has been siloed from policy conversations and broader conceptions of the internet. It is estimated that three billion people around the world play digital games, with a projected market value of more than \$300 billion by 2026. Gamers have for years discussed games and organized communities on internet forums, generated a market for platforms like Discord and Twitch, and, of course, used traditional social media to promote, discuss, and share information. In this sense, the gaming ecosystem is a significant, yet under-examined part of the existing digital world. It is also the ecosystem through which the bulk of the immersive or XR technology and content is built, and is actively experimenting with applications of distributed technologies and AI.

The gaming ecosystem is also global in scope and mirrors many of the broader debates over questions of critical technology, investment, ownership, and norms. Many of the world’s largest gaming companies are headquartered in the United States and Europe, with significant players also found in Japan and South Korea. Many of these dominant players have received significant investment from Chinese and Saudi Arabian government-backed firms, with both countries placing significant emphasis on developing ownership stakes in foreign gaming companies and increasing the reach of their own industries in the lucrative market. If these technologies become core components of the future web, then understanding the impact such investments may have on market incentives, content, product, and trust and safety practices will be important.

Further, because gaming companies have long had to work across mixed media contexts, including audio, video, and text, there are lessons to be learned from the industry’s successful and less successful

approaches to content moderation, trust and safety, and product design, as such elements become standard across the existing digital ecosystem.

Understanding what that gaming ecosystem is, how it operates, and who the key players are within it is, therefore, important to conversations about what will be needed to ensure that the digital spaces where people interact in the future are safe and manageable, and that those conversations reflect the full scope the digital world is likely to encompass.

A CATEGORIZATION OF COMPANIES IN THE GAMING ECOSYSTEM

The following section illuminates the wide range of sectors or functions that comprise the gaming industry, ranging from gaming content itself—and those creating, shaping, licensing, and selling it—to the devices used to access and engage with it. This does not constitute a comprehensive or exhaustive list of companies or players. Rather, it is an attempt to make the industry and its related ecosystem clearer to those just beginning to engage with it by explaining its core components through an illustrative categorization exercise. The following describes companies and tools across the categories of: hardware, content, game developers, development technology, storefronts, publishers, and related ecosystems.

A FRAMEWORK FOR UNDERSTANDING THE GAMING ECOSYSTEM

This chart summarizes the “Categorization of Companies in the Gaming Ecosystem” section of this annex. The companies, products, games, and platforms listed here are intended to be illustrative, and not exhaustive.

CATEGORY	EXAMPLES			
HARDWARE	Playstation console Apple iPhone	Dell desktop Meta Quest	Snap Spectacles Nintendo Switch	Xbox Console
CONTENT	<i>Call of Duty</i> <i>Candy Crush</i>	<i>Beat Saber</i> <i>Sims</i>	<i>Roblox</i> <i>Fortnite</i>	<i>Pokémon Go</i>
GAME DEVELOPERS	SuperGiant Naughty Dog	Rockstar Games	Niantic	Electronic Arts
DEVELOPMENT TECHNOLOGY	Unity	Unreal Engine	Frostbite Engine	GameMaker
PUBLISHERS	Activision-Blizzard Ubisoft	Sony	Microsoft	Nintendo
STOREFRONTS	Steam Google Play	Apple App Store Meta Quest Store	PlayStation Store	Xbox Store
RELATED ECOSYSTEMS	Twitch Discord	Reddit NeoGAF	<i>Minecraft</i> Forum Twitter	YouTube

HARDWARE

Games are played on hardware that includes both specialized machines, i.e., gaming consoles (e.g., Microsoft Xbox or Sony PlayStation) and headsets (e.g., Meta’s Meta Quest or Sony PlayStation VR), as well as multipurpose computing devices, such as laptops and desktop computers (e.g., PC, Mac), tablets (e.g., Apple iPad or Samsung Galaxy), and smartphones (e.g., Google Pixel or Apple iPhone). Until recently, most games needed to be installed from a physical disc or cartridge or downloaded onto these devices in order to be playable. Increasingly, games are also, or instead, accessed and played directly through an internet connection. Dedicated “cloud gaming” services—such as Microsoft’s Xbox Cloud Gaming, Google Stadia (now defunct), and Amazon Luna—are gaining in popularity. An increasing number of games require the player to be connected online for the game to run, even if the game has been downloaded or otherwise installed onto the player’s device. In the context of extended-reality technologies, in addition to the aforementioned virtual-reality headsets, hardware like Snap Spectacles or Meta’s RayBan Stories represent a developing market of new products that expand the range of ways in which related content can be accessed.

As measured by market share and stated preference of gamers, the most popular way to play games is on mobile phones, with consoles a distant second, closely followed by personal computers. While virtual-reality (VR) game revenues are steadily rising, (hitting \$1.4 billion in 2021) VR is still a fairly marginal piece of the gaming market, with only 7 percent of players identifying VR equipment as a preferred device on which to play games. Still, many people access content across a range of devices, meaning they may play games on their PC, phone, and console in parallel.

To give a brief sense of the scale of the gaming hardware market, it is currently valued at \$39.3 billion, and dominated by the “big three”: Microsoft, Sony, and Nintendo. VR games are largely played on VR headsets, although slimmer “goggle” devices are also available. Meta leads the way in VR headset sales with 75 percent of the market share, generating \$5.2 billion in sales between 2016 and 2021. Sony makes a headset called PlayStation VR, which requires a PlayStation console to operate. It holds 5 percent of the market share (\$5.2 billion between 2016–2021). The HTC Vive Pro 2 holds 3 percent, with \$2.6 billion in revenue.

CONTENT

PURCHASED AND DOWNLOADED GAMES

As discussed above, content in this ecosystem consists of games and interactive formats that are downloaded or accessed through physical media and played on a console, computer, mobile phone, headset or other device. That might look like a well-known game like *Call of Duty* played on an Xbox or PC, a game like *Candy Crush* played on a mobile phone, or even a game like *Pokémon Go* leveraging the augmented-reality capabilities of most smartphones. Increasingly it might include someone playing a game like *Beat Saber* on Meta’s Oculus headset. Games, like movies, span a wide spectrum of what is called “gameplay,” meaning everything from action-oriented games (like a “first-person shooter,” “fighting,” or “battle royale”) to more narrative adventures, puzzles, strategy, sports, or reflex- and physics-based games. There are also role-playing games, including well-known massively multiplayer online role-playing game (MMOs) like *World of Warcraft* or *Fortnite*.

USER-GENERATED CONTENT

Some of these MMOs constitute some of the most popular games right now, and include features that, in some ways, merit their own unique category due the ways in which users generate portions of the key content and interactions themselves. These “user-generated content” (UGC) game platforms provide users with tools, templates, and the underlying code required to make and share their own games. This includes Roblox Corporation’s *Roblox* (58.8 million daily active users worldwide in the fourth quarter (Q4) of 2022), Epic

Games' *Fortnite* (250 million registered users worldwide as of 2020), and Microsoft/Mojang's *Minecraft* (141 million active players worldwide as of 2021). The level of engagement, personal, creative, and commercial investment, and blurring of virtual and "real" worlds unfolding in these games has led several pundits and scholars to describe *Roblox* and other UGC-driven open-world games as early versions of the "metaverse."

UGC PLATFORM EXAMPLE: ROBLOX

Roblox is one of the most popular UGC platforms in the world right now. By the end of 2022, *Roblox* claimed fifty-nine million daily active users, 25 percent of whom were under the age of nine years, and an additional 29 percent were aged 9–12. A study conducted in 2020 by market-research firm Dubit found that, in the United States, 51 percent of all kids aged 9–12 had played *Roblox* in the last week. *Roblox* is a UGC platform that lets you make your own games or "experiences," which can then be shared and played by other players. Creators can also make items, costumes and other in-game assets, and share or even sell them to other players through an in-game marketplace. To date, more than 90 percent of content in the *Roblox* marketplace has been generated by player-creators, with developers and creators publishing more than fifteen thousand new experiences every day. In 2022, *Roblox* generated \$2.2 billion in revenue, some of which it "pays out" to creators who made purchasable items for the *Roblox* market. As on the *Fortnite* platform, UGC companies work with development studios to create content within the *Roblox* platform, some of which is promotional content (e.g., product placement via branded items) or immersive advertising (entire levels based on a brand or product). These applications have been critiqued as unfair or deceptive advertising. For example, consumer-advocacy groups took issue with a Walmart-sponsored "event" in *Roblox* in September 2020, involving a "Walmart Land" playable *Roblox* experience, that advocates claimed wasn't clearly labeled as an advertisement. How advertising and the creator economy emerge as revenue models within the gaming industry will be key to understanding how risk, threats, and opportunities will also emerge in that sector.

CONTENT CREATORS

Content creators in the game ecosystem include individuals, organizations, and companies ("studios") that develop games, experiences, and other immersive content as a creative industry activity (game producers or developers); as well as players who produce UGC games and assets for themselves and others to play with on an existing UGC game platform (UGC creators). Although these two groups are often differentiated as "professional" versus "amateur," there are exceptions such as professional UGC creators, and overlaps like hobbyist game developers. While some game content is now made by users, the games market (i.e., games sold and played) continues to be dominated by game developers.

GAME DEVELOPERS

Historically, studios have specialized in games played on a particular device or type of device: (e.g., mobile, computer, VR, one or more consoles). Successful games would then be "ported" to multiple consoles and/or devices (i.e., the programming code would be adapted to allow the game to be played on another device). Today, studios are increasingly making multiplatform games that can be played across platforms and devices (including VR headsets) from the outset, or shortly after release. Game-developer studios are often a complex web of companies that sit across multiple verticals of the gaming-industry ecosystem. This is further complicated by a recent trend of consolidation, with many successful smaller developers acquired by major

studios, and even major gaming companies merging into one another (Microsoft’s attempted [acquisition](#) of Activision-Blizzard is a contentious current example). Across that ecosystem, development of content happens through some combination of the below.

Indie developers/studios: Independently owned studios, most of which are not exclusively tied to a major publisher or distributor. In some instances, “indie devs” come up with an original concept or prototype, which is then pitched to a publisher that provides resources to develop the game. In others, indie devs might bring the game to beta before a publisher becomes involved. In still other instances, they develop the full game, and a publishing deal is then made to help with distribution, marketing, and (often) adapting the game to different platforms. Lastly, indie devs can also self-publish and market their games, distributing them through Steam or other major platform “storefronts” (see below, e.g., Apple Store, Oculus App Lab and Store) or online (e.g., itch.io). There are a shrinking number of indie developers as the major players increasingly acquire these studios. It is worth noting that several key developers in VR, augmented reality (AR), and immersive games are mid-tier companies that describe themselves as indie.

- ▶ **Examples:** Supergiant Games (*Hades*, *Bastion*, *Transistor*), Deck Nine, Extremely OK Games, Innersloth. And now acquired Ape (*Stardew Valley*) and Mojang (*Minecraft*)

Major industry and in-house studios: Large studios dominate the games industry and market. They develop the bulk of the biggest, most popular, and most lucrative (“blockbuster”) titles, and they publish and license games made by other studios. Over the past twenty years, the industry trend has been consolidating, with what are often referred to as triple-A studios acquiring indie and mid-tier studios, funding subsidiaries, exclusive and nonexclusive publishing contracts, etc. [Triple-A studios](#) commit large teams (many with specialized skills) to the development of high-caliber games with large production and marketing budgets, designed and targeted to sell: they offer high risk for high return. There are also a number of [in-house game developers/studios](#) owned by large media and other conglomerates that form part of this better-funded ecosystem. For example, Disney Mobile makes tie-in games for iOS and Android featuring Disney characters, while Niantic develops its own games like *Pokémon Go*.

- ▶ **Examples:** Rockstar Games (*Grand Theft Auto*), Naughty Dog (*The Last of Us*), Electronic Arts (*Madden*, *FIFA*), Activision-Blizzard (*Call of Duty*, *World of Warcraft*, *Diablo*), PlayStation Studios (owns Naughty Dog), Microsoft Game Studios, Ubisoft (*Assassin’s Creed*), Niantic (*Pokémon Go*)

Outsourced labor: A growing portion of [development processes](#) are [outsourced](#) to companies located in countries with poor labor standards and regulations. This is especially common among the triple-A studios, but many mid-sized studios also outsource components of the development process. As more elements of game design and development are shifted to AI and procedural generation, [outsourcing](#) is not so much reduced as shifting to “conditioning” work, a term used to describe the fine-tuning or tweaking of AI-generated results done by human workers. There are sizable communities of game developers that work with these larger companies in Russia and China, in particular, with further concentrations across Eastern Europe and Asia.

DEVELOPMENT TECHNOLOGY

These are the companies that provide the technology that content developers use to create games, many of which are made using software systems called “game engines.” A game engine provides the computer code framework for developing a game, which includes libraries of resources used by computer programs (like prewritten code), and various tools and assets (i.e., a representation of an item that appears in the game). Game engines typically provide the core functionality of the game—for example, they can supply the program that simulates “physics” (e.g., gravity) within the game environment. They are akin to Photoshop for

image creation, or Microsoft Word for written documents: providing templates and customization options that users choose from and build on to create new works. Multiplatform game engines such as Unity are heralded as having “democratized” game development by making it more accessible to beginners and to those without advanced programming skills.

There are a few dominant game engines, though most of the largest and most profitable game studios use their own, custom game engines. This includes the following.

- ▶ **Unity (developed by Unity Technologies):** Currently one of the most popular game engines in the world, Unity is particularly valued for its user friendliness, flexibility, and community support, including vast sources of freely available documentation. Unity Technologies offers a tiered-licensing system that makes the engine accessible to small and large studios, as well as individuals, and provides a free licensing program for nonprofits and educational institutions. In 2018, the company claimed that half of all games were built using Unity, accounting for approximately \$35 billion of the 137.9-billion game industry. Of note, Unity has also expanded its technology for use in the auto, film, and architecture industries, and is particularly popular for mobile-game development.
- ▶ **Unreal Engine (developed by Epic Games):** Though fewer games are made using Unreal Engine, it is nonetheless a popular and powerful engine. Particularly noted for its advanced graphics capabilities and available tooling for large-scale projects, Unreal is especially popular for higher-budget projects or those requiring high visual fidelity. Epic also provides full access to Unreal’s source code, making it not only adaptable but easier to debug. Though less user friendly than other engines like Unity, the introduction of its visual-scripting language Blueprints, as well as friendly pricing structures for small studios, has made Unreal Engine increasingly popular. The company developing this engine, Epic Games, also uses it in house, including for its massively successful game, *Fortnite*.
- ▶ **Proprietary game engines:** Most triple-A studios use their custom-built game engines, including Nintendo, Rockstar (Rockstar Advanced Game Engine, RAGE), and Electronic Arts (EA/Frostbite). These are mostly made in house or commissioned for exclusive use, and are proprietary. A decreasing number of indie studios build their own game engines due to the high cost involved, paired with the prevalence and ease of use of popular game engines like Unity. It’s not clear the extent to which proprietary engines will be used for VR experiences. For example, EA announced in 2016 that it was developing Frostbite Labs to create triple-A virtual worlds and assets, but its recent high-profile VR releases (e.g., *Medal of Honor: Above and Beyond*, 2020) were developed by mid-sized studios (e.g., Respawn Entertainment) using multiplatform engines (e.g., Unreal).

Godot and GameMaker are additional interesting examples to consider. Increasingly, young people are finding their way into game development through Scratch and other visual-programming languages.

While not exactly the same as a game engine, game platforms such as *Roblox* and *Minecraft* provide a similar, albeit much simpler, function—through the provision of templates, tools, and the core underlying programming that players can then use to create their own UGC games, experiences, and assets (such as items that can be used or worn in game, objects that appear in a game environment, etc.). These UGC game platforms working to provide “low-code” or “no-code” tools are mostly used by amateur and pro-am users, but, in some cases, are used by developers to create games, and used to create “immersive advertising” for or by third-party companies.

PUBLISHERS

Game publishers play a similar role in the games industry to that producers do in film. They finance games that are developed either in house or externally (or in collaboration between the two). They can own multiple smaller and mid-tier studios, and/or have exclusive (or nonexclusive) rights to the games that these studios develop. Publishers can get involved at various stages in a game's development, including at the outset (e.g., games developed in house, game concepts that are pitched by indie developers, games developed by subsidiaries), midway, or after the game is complete (or near complete). They often have a say in what goes into the game because they provide critical resources, including market research, creative teams, additional developers to "finalize" the game or enhance its production value, sound designers, etc.

Large publishers have in-house resources that they mobilize across both internal and external development projects, some of which are highly specialized, such as [Ubisoft's Performance Capture Studio](#). They also usually take charge of promoting and distributing the game (or paying for distribution), licensing and "localization" (translating or adapting a game for foreign markets), and handling various other elements aimed at helping a game succeed in finding its audience (e.g., showcasing the game at industry conventions or fan expos).

Many major game publishers are also game developers, with in-house studios as well as game distributors. Five publishers also make hardware on which games are played: Sony PlayStation consoles/VR headsets, Microsoft Xbox consoles, Apple iPhones/iPads, Nintendo Switch consoles, and Google phone. All of these companies are large multisector conglomerates, illustrating a global trend in the games industry toward consolidation as major industry players buy up indie and mid-sized game-development studios. This mirrors and overlaps with what has happened within social media and other tech platforms (e.g., Meta acquiring Instagram, Google acquiring YouTube). Indeed, some of these same players are driving consolidation in the game industry, including Meta and Google.

STOREFRONTS

Once content is published, it is distributed through storefronts, some of which are specific to the hardware that will be used to access that content (VR headsets, phones, etc.). This could mean downloading a game like *Call of Duty* through the Xbox Store, *Candy Crush* or *Pokémon Go* from the Google Play store, or *Beat Saber* through the MetaQuest Store. Games are now mostly (83 percent) sold in [digital format](#), though there is still a small (17 percent) market for [physical copies](#) (i.e., discs or cartridges played on a console or computer). Below are additional examples of dominant storefronts broken down by hardware type.

- ▶ **PC/desktop game-distribution platforms:** Steam is the market leader in PC-game sales and indie-game distribution. The company reported [132 million monthly active users](#) in 2021 and [released 10,963 games](#) in 2022 alone. Other desktop distribution platforms include Epic Games Store (sixty-two million active monthly users in [2022](#)), GOG.com, Apple Store, Microsoft Store, and Google Play.
- ▶ **Mobile-game distribution platforms:** Apple Store, Google Play, Samsung Galaxy, Amazon Appstore, Microsoft Store, Huawei AppGallery.
- ▶ **Console-specific game distribution systems:** Xbox Store, PlayStation Store, Nintendo eShop, Meta Quest Store, Steam Store (for Steam Deck).
- ▶ **Cross-platform integration:** Many storefronts have their own apps that allow users to access the store to purchase games and other limited features through their mobile phones for use on multiple devices (Xbox Store, Steam, etc.).

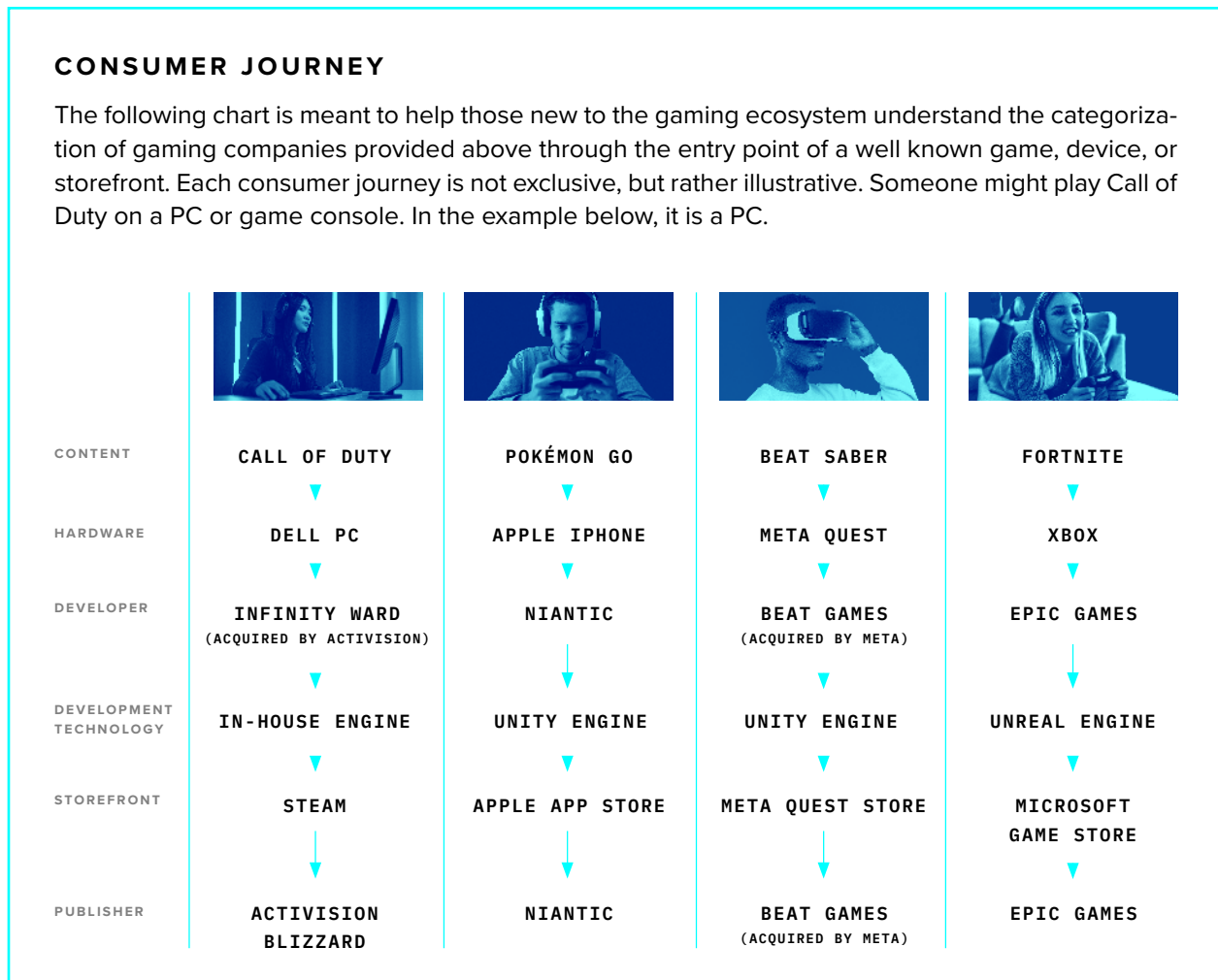
On all of the major console-based gaming platforms (e.g., Microsoft Xbox, Sony PlayStation, Nintendo), special access to games and other features, including online multiplayer, are bundled and sold in the form of subscription-based annual or monthly memberships through their proprietary virtual storefronts. The most prominent example is the Xbox Game Pass, a cross-platform tiered subscription service that is required for accessing cloud gaming and for playing online multiplayer games and “modes.” Microsoft claims to have twenty-five million Game Pass members worldwide.

In this way, digital game distribution can be dominated by companies that make the hardware on which games are played, as well as the games people are playing (Microsoft Store on Xbox, Sony’s PlayStation Store, and the Nintendo eShop). However, Valve’s indie-game-focused Steam distribution service and storefront is an example of a more open platform that provides a key distribution mechanism across this ecosystem. Vast libraries of games are available for purchase (or free download) through these storefronts, including ones the storefront’s parent company has developed, as well as those made available by third-party publishers or developers (often for a fee and commission on sales).

Game-distribution storefronts that are embedded in associated gaming platforms serve as more than just a way to buy games. Microsoft’s Xbox and Sony’s PlayStation, for example, seek to create a sense of continuity across games and devices, and to support a community or social network among players. Each of them has social-connectivity aspects that allow users to engage with one another on the platforms in ways that may look and feel like traditional social media, with profiles, friends, and other related features.

CONSUMER JOURNEY

The following chart is meant to help those new to the gaming ecosystem understand the categorization of gaming companies provided above through the entry point of a well known game, device, or storefront. Each consumer journey is not exclusive, but rather illustrative. Someone might play Call of Duty on a PC or game console. In the example below, it is a PC.



ADJACENT COMMUNITIES AND PLATFORMS

The social aspect of gaming is not limited to these gaming-platform accounts. Game-player communities have long included numerous very active, early adopters of online tools and forums. Many of these communities are integrated into game development (invited to playtest, provide early feedback, etc.) and the promotion of new or upcoming games. As a tech-savvy group, these communities can be found across the internet: online gamers make up for a significant part of social online spaces such as Reddit, and routinely gather the biggest audiences on social media sites (e.g., PewDiePie on YouTube). These communities play an important enough role in the broader online ecosystem that it is important to understand the main platforms through which they engage, communicate, and collaborate.

Gaming communities use these platforms to discuss approaches to games, watch others play games, and share opinions. Some communities have also leveraged these platforms to coordinate harassment and real-world harm. In this way, issues of trust and safety are significantly affected by communities on these adjacent platforms, as trends, coordinated harassment, and workarounds for game content controls and safety mechanisms may be shared and facilitated in these spaces, as well as in chat functions within games themselves. One of the most insidious examples of this was the [#Gamergate campaign](#), which involved gamers coordinating targeted and real-world harassment against female gamers through platforms like 4chan, 8chan, Reddit, and, eventually, Twitter.

Some of the key platforms and forums developed for gamers have since become mainstream platforms far beyond games. This includes platforms like server-based [Discord](#) and live-streaming platform [Twitch](#). In other cases, gamers leverage existing platforms in unique ways, as is the case with Reddit's heavily used [r/gaming](#) or [r/games](#) subreddits. Gamers also gather on a range of gaming-specific message boards. Additional information on these platforms can be found below.

- ▶ **Platforms developed for gamers that are finding wider use:** Discord is a US-based platform centered on voice communication (live, recorded messages) and posting (text, images, etc.) in private or community chat rooms called “servers,” each of which is managed or moderated by different users. It has more than [150 million active monthly users](#), and game-related servers are still some of its most popular (e.g., Blox Fruits, a *Roblox*-focused server, has one million users). Twitch is a US-based live-streaming and video-on-demand service focused primarily on video games, including electronic sports (esports), as well as other media topics and genres. It was [purchased by Amazon](#) for close to \$1 billion in 2014. It has more than [7.4 million active streamers](#), and is still home to many video-game players who monetize their playing through Twitch streaming channels. Gamers also use and engage through Google's YouTube.
- ▶ **Gaming-specific messaging boards:** Gamers gather in a number of forums, ranging from those hosted by games themselves (*Minecraft* Forum) to storefronts like Steam, as well as a number of gaming-specific messaging boards that vary greatly in their moderation policies. For example, messaging board ResetEra split from messaging board NeoGAF after a controversy over the behavior of NeoGAF's founder and a desire for more stringent moderation policies.
- ▶ **Traditional and niche social media:** Gaming communities also communicate and coordinate through major social media platforms such as Meta, Twitter, and Reddit. They also have organized using less prominent platforms like 4Chan or 8chan, which long served as a center of gravity for gamers, with a number of popular boards focused on games, game developers, and players. The site has been subject to endless controversies for its alleged association with radical and extremist political movements, and is reportedly closely associated with the “toxic gaming culture” widely covered in the press.

CURRENT GAMING BUSINESS MODELS

The integration of online features in game hardware and software has led to a number of new business models focused on content monetization, user-data collection, and advertising. Most players of multiplayer games expect new content and storyline developments to emerge over time, especially if their gameplay extends over several months, or even years. This ties into larger industry trends to expand on a game's original content through the development and release of add-ons, expansions, extra levels or characters, and other forms of downloadable content (DLC). This supplementary content is most often made available for purchase as a monetization strategy, which has been significantly facilitated by the rise of digital distribution platforms. According to Unity's Gaming Report 2023, the lifespan of existing mobile games was extended by 33 percent in 2022 due, in part, to regular updates. Notably, larger studios are more likely (and able) than indie devs to update games for more than six months following release (84 percent vs. 55 percent).

The ongoing release of additional content is a central component of “freemium” and “free-to-play” revenue models and other micro-transaction-based monetization strategies, and part of the broader shift toward publishing games as a “service” rather than an end-user product. A key example here is *Fortnite*. According to a survey conducted in 2020 (before the game surged in popularity during the pandemic), *Fortnite* players had spent an average of \$102.52 on in-game purchases. Overall, 77 percent had spent money on in-game items, avatar clothing, and other micro-transaction purchases. The shift to free-to-play and other micro-transaction-based business models is changing how games are designed and the types of player experiences that are prioritized in game reward systems. It is also worth noting that video-game companies, for these reasons, have found themselves at the forefront of challenging app stores on their fee structures, a notable example being the 2020 case Epic Games v. Apple.

Much like other internet-based platforms, games are designed to be “sticky,” to keep people entertained and engaged. That means they can also incentivize predatory monetization hooks, such as loot boxes, prompting players to spend real money in games for virtual (and uncertain) content. Public outrage over loot boxes' similarity to online gambling, and the perception that they are particularly targeted to children, has resulted in government authorities and industry associations exploring outright bans, disclosure requirements, and other regulations. As a result, the practice has significantly decreased, but is a useful reminder of incentives at play.

Because people either purchase a game outright or purchase additional features or experiences within a game, micro-targeted ad-based revenue is a less prominent feature in the ecosystem to date than on other internet-based platforms (mobile games, in particular, make use of in-app ads for other games). However, as immersive options grow in this industry, a point of significant debate is whether companies are likely to turn to this data-collection-intensive approach, with so many additional sources of biometric and other interaction data available and likely interest from advertisers in adjacent industries.

There has also been a fair amount of interest in a burgeoning Web3-based gaming industry, partially due to the new and innovative business models promoted as beneficial to both gamers and game developers. These models include play-to-earn, non-fungible tokens, decentralized autonomous organizations, and subscription-based models. The most prominent example was a game called Axie-infinity, created by a Vietnam-based company, which was met with a great deal of publicity as it seemed to break through as a crypto-based game enabling players to earn money by playing. About a year later, the acclaimed game was hacked and just about completely collapsed, leading to a tempering of enthusiasm about Web3 gaming potential.

MARKET LEADERS IN GAMING AND THEIR COUNTRIES OF ORIGIN

While there are countless companies working across each of the categories above, a number of the largest industry players sit across multiple categories. In this regard, the gaming ecosystem is not dissimilar from traditional Web2 industry monopolies. One key difference is the geographic diversity of the largest industry actors and their owners.

Currently, the top game developers in the world, in terms of market value, are also publishers. This includes companies in the United States, Japan, France, Poland, and Sweden. As of April 2023, their respective market values were: Activision Blizzard (\$67.06 billion, United States), Nintendo (\$47.8 billion, Japan), Electronic Arts (\$35.12 billion, United States), Roblox Corporation (\$24.55 billion, United States), Take-Two Interactive Software (which owns Rockstar and 2K; \$20.98 billion, United States), Square Enix (\$5.72 billion, Japan), Embracer Group AB (formerly THQ Nordic/Nordic Games; \$5.12 billion, Sweden), Ubisoft (\$3.04 billion, France), and CD Projekt (\$2.64 billion, Poland).

The top gaming developer in the world, Activision Blizzard, owns King.com Ltd., the makers of *Candy Crush*, which was the highest revenue-earning mobile game in 2022 (NDP Group, 2023) and one of the most popular (i.e., most downloaded) games of all time. As of January 2023, Activision's market value reflected its \$75-billion proposed acquisition by Microsoft, which, at the time of writing, is being challenged by regulators in the United States, European Union (EU), and United Kingdom (UK) on the basis of Microsoft's existing dominance in game distribution and hardware.

Legacy gaming companies also play a prominent role in the development of immersive content, VR, and AR games (combined under the term XR). A recent study of professional game developers found that 38 percent were involved in VR/AR game development, and 23 percent said that the platforms that most interested them were VR headsets, especially Meta Quest and Sony's PSVR2. Both indie and triple-A studios are leaders in XR game development. One of the most downloaded VR games to date, *Beat Saber*, was developed by Czech indie studio Beat Games (which has since been acquired by Meta).

Most of the top game companies worldwide in terms of revenue are game publishers with significant overlap with hardware producers and some developers. This includes companies in China, Japan, the United States and Singapore. In 2021, the top ten were: Tencent (\$32.8 billion, China), Sony (\$18.2 billion, Japan), Apple (\$15.3 billion, United States), Microsoft (\$12.9 billion, United States), Google (\$11 billion, United States), NetEase (internet and video-game company; \$9.6 billion, China), Activision Blizzard (\$8.1 billion, United States), Nintendo (\$8.1 billion, Japan), Electronic Arts (\$6.5 billion, United States), and Sea Limited (a game-development and publishing company turned tech conglomerate; \$4.3 billion, Singapore).

Note that in this space, China's Tencent has more than twice the revenue of US tech giant Apple, and almost twice the revenue of its next closest competitor, Japan's Sony. This brings home both the scale of the gaming market and the fraught geopolitical implications of the industry's geographic distribution.

GLOBAL INVESTMENTS IN IMMERSIVE TECHNOLOGIES AND GAMING PLATFORMS

Just as technologies flow and originate from around the world, so do investments in gaming and technology platforms. Chinese tech giant Tencent is a leader in gaming and a major investor in gaming platforms globally. These investments are notable because many of the gaming companies in which Tencent and other Chinese firms have invested are based in the United States, European countries, Japan, and other US ally and partner nations. Tencent's investments in the gaming space internationally span ownership of Riot Games (United

States), Funcom (Norway), Sharkmob (Sweden), and Leyou Technologies Group (Hong Kong); and stakes in Remedy (about 5 percent, Finland), Epic Games (about 40 percent, United States), Activision-Blizzard (about 5 percent, United States, amid Microsoft's attempted acquisition), Ubisoft (about 10 percent, France), Krafton (about 10 percent, South Korea), and Supercell (about 70 percent, Finland), among others.

In 2021, UK-headquartered gaming firm Sumo sold to Tencent for \$1.26 billion. Tencent had previously invested \$150 million in Reddit in 2019. China's esports market grew by 14 percent in revenue from 2020 to 2021, even though the Chinese government imposed a tough regulatory environment, and the biggest esports streaming apps in China currently boasts millions of users, including Huya (about 30.46 million), Douyu Live (26.25 million), and Huajiao (9.97 million). Chinese tech giant Alibaba made a push into gaming in 2020, but has not focused much on the area since.

Saudi investors likewise are an increasingly major player in the international gaming-platform market. In April 2023, the Saudi Arabian government announced a strategy to invest \$38 billion in making Saudi Arabia a global video-game industry hub. Already, the Saudi sovereign wealth fund holds an 8.26-percent investment in Nintendo, runs the Savvy Games Group looking to establish two hundred and fifty gaming firms in Saudi Arabia, and holds stock in Activision Blizzard, Electronic Arts, and Take-Two Interactive.

For both China and Saudi Arabia (as well as other countries), investments—and strategic, geopolitical, and profit opportunities—also lie with the infrastructure that supports gaming platforms and services. For instance, Tencent is a major cloud provider. As the cloud-gaming market in China grows rapidly, Tencent is positioned to benefit greatly from both stake in gaming companies and control of cloud infrastructure that can support delivery to their users. Ironically, Saudi Arabia's large investment in gaming companies and platforms could also benefit a different Chinese firm, the telecom Huawei, which has recently expanded its cloud-service infrastructure in the country.

These developments raise many points. While it does not receive as much media attention as social media platforms, the virtual gaming space is exploding across the world, and governments and companies with the funding, technology, and/or underlying infrastructure are taking notice. It is another realm in which countries are economically competing with one another in technology. The gaming companies receiving investment often build technologies that go far beyond gaming itself, whether designing buildings and cars, filming movies, or building simulations for militaries.

Exactly what interest each country has in these companies may vary, but it is clear they believe that the gaming industry and the growing use of its technology are strategically important enough to incentivize significant and coordinated investment. The broader trust and safety risks associated with potential foreign engagement with gaming platforms (from ownership, to code development, to infiltration of platforms for other purposes) relate to broader debates about internet governance as well.¹

IMPLICATIONS FOR FURTHER STUDY AND EXPLORATION

The intention of this annex has been to provide a basic introduction to the gaming ecosystem, including the key players, companies, technologies, nations of origin, and sources of investment and profit. The remaining sections will explore the implications of these components, identifying questions requiring greater examination and areas for investment and policy attention.

¹ For a deeper analysis of this topic, see *Annex 6: Learning from Cybersecurity, Preparing for Generative AI*.

TRUST AND SAFETY

In some ways, the gaming world has pioneered new types of harm, eventually replicated in other digital spaces. It has also pioneered innovations to address those harms and proactively build better interactions. Because the gaming world has been siloed from traditional policy conversations and trust and safety communities, this remains an under-examined area of practical study. This is even more so as elements of gaming and legacy interactive media begin to merge, particularly when it comes to immersive contexts. There are a number of particular elements of this double-edged gaming sword to examine.

LEARNING FROM GAMING'S CHALLENGES AND UNIQUE PERSPECTIVE

Games have long existed as multimedia interactive spaces that commingle audio, video, and text components as a key feature of the technology. Importantly, many of these multimedia interactions occur in real time. This differs from traditional social media, which have been dominated by text-based, asynchronous platforms with slower adoption of audio, video, and real-time interactions, and less familiarity with managing the risks involved in these contexts. As immersive technologies become more pervasive, they are not only likely to stem from the gaming ecosystem, but will replicate many of the dynamics with which games have long grappled.

Another unique element of games is its intentional inclusion of children, including those younger than thirteen, and adults. Whereas most traditional social media platforms work to prevent their products and platforms from being used by children, the gaming industry regularly builds for, and markets to, them. Many games are published with a rating (more on that in the section on regulation), which means game developers often declare an intended audience before releasing a game for play. As debates grow over the impact of technology on children, and the best approaches to keeping them safe online, lessons from the gaming world are worthy of closer examination.

There is also a [well-documented](#) history of harassment, hate speech, sexism, and racism in the gaming world. In more well-known cases like the [Gamergate](#) scandal, this harassment can move from gaming-community spaces to mainstream social media, and break into violence and harassment in the physical world. [Some credit](#) the methods of harassment honed through the Gamergate process with informing the approach and success of those who used the internet to spread false narratives about the US election in 2020, and to organize the January 6, 2021, insurrection. Regardless of where one comes down on those linkages, for those seeking to address toxicity and harassment in digital spaces, it is unquestionable that some of the mainstream trends we struggle to address in today's digital world have roots in the gaming world. As these linkages will continue, we would be wise to treat the gaming ecosystem as part of the broader information environment moving forward.

LEARNING FROM GAMING'S INNOVATIONS AND APPROACHES

The flipside of these challenges and unique gaming features is that the gaming industry's approach to policy, trust and safety, tooling, and design differs significantly from that of traditional social media. Better understanding this may illuminate new tools and approaches for our future digital world. It also may provide lessons on what does not work in solving problems that may be new, or simply new to certain sectors.

One of gaming's unique areas for further exploration is its focus on designing intended interactions. As opposed to traditional social media, which often focus on a concept of neutrality, most games are an exercise in trying to create specific kinds of interactions for a player with content, or with other people, within a highly controlled environment. This does not mean those intended interactions are normative—at the most basic level, a game defines how your character can move through a story, challenge, or series of skills. An intended interaction may be to kill an enemy, or to work with other players to open a door.

However, in recent years, some in the gaming industry have added a normative frame to game development, pioneering “prosocial” approaches—more intentional and proactive design methods that preemptively shape and encourage healthy and inclusive play patterns. These methods pull from best practices in design, psychology, sociology, human factors, and more, as well as case studies from earlier multiplayer games. Those at associations and organizations like the [Fair Play Alliance](#) are experimenting with these approaches as a complement to more traditional moderation tools, with the goal of reducing net harms and improving recidivism rates.

As some of the early companies building immersive content, gaming companies have also begun to leverage player dynamics and safety by design in the conception and construction of virtual worlds and experiences. With the increasing popularity of VR games and applications, companies are focusing on developing new safety features to protect users in these immersive environments. These features can include limiting play-time, providing warnings for potential motion sickness, allowing users to control who can interact with them in any particular gaming space, developing stronger tooling to support real-time monitoring in dynamic-video and audio-based user environments, and ensuring that users are aware of their physical surroundings while playing. Efforts to improve the ability to provide real-time monitoring in privacy-respecting and less data-intensive ways will have applications for numerous industries.

There are also long-standing challenges that gaming companies have prioritized solving that could have application in a broader trust and safety context. One example is anti-cheating policies, enforcement, and tooling. In online games, some players seek to use third-party software or other hacks to augment aspects of a game to their advantage. If this behavior reaches a critical threshold, online games can become almost unplayable for those not cheating. As a result, game developers have [created a number of tools](#) to make cheating harder, disincentivized, or punished. Some of these techniques, whether throwing cheaters on servers full of only other cheaters, triggering insurmountable challenges or a massive increase in difficulty levels as a result of certain behaviors, or effectively instituting timeouts, have corollaries in traditional interactive spaces. It is worth examining the wide arrange of approaches gaming companies use to shift the incentives of certain behaviors in their ecosystems.

It is also worth noting that there is an established history within the gaming ecosystem of various forms of community moderation. Whether *League of Legends* Players Tribunal, Reddit gaming threads, Discord channels, or any of the countless gaming message boards, the gaming world has leaned into the idea of enabling unique rules and norms for unique spaces, set and enforced by communities. Better understanding the mechanisms, benefits, and drawbacks of these approaches could have utility in other digital contexts.

ILLUMINATING EXISTING GAMING CONTENT MODERATION AND TRUST AND SAFETY LAYERS

Just as in other digital sectors, the term “trust and safety” is relatively new, and encompasses a wide range of teams, functions and tools in gaming. In recent years, there has been a significant rise in the focus of gaming companies on addressing problematic behavior and harms within games, leading some to [build new in-house trust and safety teams](#) and recruit established trust and safety professionals from other digital industries to lead them. One of the challenges of applying this lens to the gaming industry, however, is just how spread out and disconnected from one another existing content policies, security requirements, and trust and safety mechanisms are throughout the tech stack described in the first part of this paper.

For example, storefronts like app stores have privacy requirements that apps must meet in order to be approved for inclusion; game studios have content standards and rules they require their developers to incorporate; game developers actively design features of games to minimize certain harassing or cheating

behavior; and interactive games systems allow users to report bad behavior or inappropriate content directly within a game. Each of these policies may be set and implemented by a different company, with players largely unaware of them. More can be done to map and understand these interlocking layers as part of the trust and safety ecosystem.

These are just a few of the areas worthy of exploration and study. We recommend additional work to better connect the gaming ecosystem to other media and information spaces, exchange approaches and information between their trust and safety, development, and other communities, and explore areas of vulnerabilities and opportunities.

REGULATION AND GOVERNANCE OF THE GAMING SECTOR

As more of the gaming ecosystem and social media-dominant digital spaces converge, questions of which regulations and oversight bodies might apply, and in which ways, is an area deserving more study and policy clarification.

Regulations affecting gaming companies include policies and laws that relate specifically to games, as well as some media regulation, consumer-protection laws, privacy and data-protection laws, laws prohibiting hate speech and/or protecting freedom of speech, laws focusing on child safety, copyright laws, and the regulation of digital services and content.

Of particular note are game rating systems, set by regional game content-classification bodies, each with their own standards and rules, mostly focused on age appropriateness. In the United States, console and some computer and mobile games are rated by the Entertainment Software Ratings Board (ESRB), an industry organization similar to the Motion Picture Association (MPA) for film. Compliance is voluntary. The EU and UK both follow the PEGI system (Pan European Game Information), which is managed by both industry and government, and enforced by governments in various participating countries. Neither system rates online interactions, although PEGI oversees a supplementary “PEGI Online” certification system for games that “commit themselves to banning inappropriate material” and interactions between players. In Japan, the Computer Entertainment Rating Organization (CERO) consists of a rating system similar to PEGI and the ESRB. South Korea has a governmental organization called the Game Rating and Administration Committee (GRAC). Other nations and regions have other systems. None provide a comprehensive review of online interactions.

For a company to release a game in each of these markets, it often must undergo review to ensure it complies with each individual standard. In a country like China, where the government sets strict content standards that it enforces itself, companies often need to create a unique version of their game to comply when a common denominator is not possible.

China, it is worth noting, is a completely unique ecosystem, with tighter regulation around nearly all aspects of game development, operation, and use. For example, children and teens are banned from online gaming during school days, and have been limited to one hour a day on non-school days. Companies looking to sell their games within China must be able to demonstrate that their games can comply with and support such restrictions, and are subject to review by the Ministry of Culture. Mapping geopolitical dynamics, and understanding which rating systems dominate game-company decision-making, where these ratings systems do and do not apply with regard to new forms of interactive media, and where other regulatory regimes might become relevant are all areas worthy of greater examination.

Another area of particular focus is privacy and data collection. While gaming companies have long collected data on their users and their interactions, because their business models have not been dominated by

hyper-personalized ad models—as has been the case in the social media ecosystem—conversations have focused less on this area of policy. However, with increasing forms of immersive and interactive technologies creating opportunities for new kinds of personal, behavioral, and biometric data, many people are concerned this will create an incentive for the dominant social media business model to become more common in gaming spaces. This presents new risks worthy of investigation and deeper study.

THE RISE AND IMPACT OF GENERATIVE AI IN GAMING

As is the case in each section of this report, the rapid application of generative AI within existing and developing digital ecosystems presents a number of areas requiring close attention. Within gaming in particular, generative AI is already playing a role in content creation, creating new questions about intellectual-property rights, as well as privacy concerns.

FINDING SIGNAL AMID NEW TECHNOLOGY HYPE CYCLES

As discussed throughout this paper, no one is certain which of the emerging connective tech trends will become dominant, pervasive, or game changing. Readers would be forgiven for dismissing all talk of the metaverse, Web3, and generative AI at various points of their respective hype cycles, but the current explosion of AI technologies brings home just how important it is to pay particular attention to the ways each of these technologies is changing existing practices and infrastructure.

For example, the massive popularity of UGC games (which straddle traditional gaming and something approximating the metaverse) raise a number of unique questions and issues, pertaining to content creation and authorship (e.g., who owns the content co-created in these games?), how monetization strategies can be expanded to players (as revenue earners, not just as consumers) as well as exploited by players (e.g., to cheat other players, or steal currency or personal data from them), how to ensure trust and safety when content is added endlessly by a diffuse and massive player population (which sometimes count in the millions), and what special protections might be needed to support the children and teen players who dominate these spaces.

Whichever technologies emerge most impactfully will inform which risks, opportunities, and resources need to be prioritized, be it augmented reality via existing devices taking off, or virtual-reality hardware becoming more widespread, or other metaverse applications, such as wearables, becoming ubiquitous. Do Amazon's investments in health tech increase its interest in device-generated data? Does that impact the content created for distribution across Amazon's platforms? Do nonfinancial applications of Web3 take hold in gaming or elsewhere? Does that impact the consolidation we are seeing within the gaming industry, or geopolitical trends of nation-state competition through this arena?

More work needs to be done to explore where these technologies bump into existing regulatory regimes and communities of practice, and how they might change those spaces.

GEOPOLITICAL IMPLICATIONS AND INFLUENCE

As discussed above, gaming and other technology platforms are clearly of interest to a growing range of nation states, and can provide them with geopolitical advantages. There is enormous profit to be made in everything from content creation to platform development to supplying the infrastructure underlying gaming platforms themselves, and significant geopolitical benefit to being market leaders. Governments could also potentially approach the companies building these gaming products and services to request data—and

could also impose content requirements on those platforms. This could range from removing content that is disfavored by the state to requiring companies entering the market to pre-censor and tailor their games to align with state narratives from the outset—a kind of economic coercion of the gaming market. The Saudi Arabian government’s \$38-billion investment plan for gaming introduces questions about these kinds of risks. Of course, China is particularly well positioned to (and often does) exert this kind of economic leverage on non-Chinese firms, given the size of its domestic market.

For example, the Chinese government has for decades cracked down on video games that do not follow the state’s narrative line. It has also punished celebrities, spokespeople, and well-known gamers for communicating unwelcome support for causes ranging from Tibet to Hong Kong pro-democracy protests. This has included pressuring gaming companies to cut off revenue streams to these influencers, or refusing to speak to certain journalists. These actions, combined with the sheer market power China presents, have also had an effect on some game content. While in the past, Chinese content requirements usually meant gaming companies would be forced to produce secondary Chinese versions of games for sale in the Chinese market, companies are increasingly developing their primary version of games in alignment with Chinese content rules, and selling those versions everywhere. Is a potential, logical extension of this a wider move to design games through a default framework of surveillance by design rather than safety by design?

Elevating these concerns, of course, are the sizable investments both China and Saudi Arabia have made in Western gaming companies. Aside from questions around their ability to influence content decisions and trust and safety approaches, some experts worry about the potential access to user data and other sensitive information. Further study is needed on investments in, and potential impact on, emerging tech and platforms, whether immersive, generative AI, or distributed.

These platforms and environments can also become targets during geopolitical tension, and even war. Aside from concerns over the cybersecurity of these systems (increasingly targeted by government and independent hackers), Microsoft’s President Brad Smith has pointed out vulnerabilities extend further, noting, for example, that Russian intelligence organizations are seeking to penetrate gaming communities to spread pro-Kremlin propaganda about the Vladimir Putin regime’s war on Ukraine.

CONCLUSION

The gaming ecosystem is massive and expanding rapidly. It is also merging with other parts of the digital ecosystem. Whether it is properly understood as part of this broader digital and information landscape will have significant impact on our ability to build the tools, approaches, and communities required to achieve a healthier version in this next phase of our connected world.

The fact of gaming’s isolation from policy communities focused on internet governance, social media, and big-tech issues has resulted in a lack of appreciation of the gaming industry’s long-standing market share, geopolitical impact, technology innovation, and connection to the rest of the information ecosystem. Because much of the emerging immersive technology is developed through the gaming ecosystem, understanding the players and how they work takes on even greater importance. This includes understanding ownership, incentives, and business models.

The gaming community, in some cases, generated—and, in other cases, was the first to experience—many of the harms, risks, and challenges everyone is grappling with today. This includes both the social impact of harassment and toxic online behaviors, as well as simply experience operating in multimedia, interactive, and real-time spaces. More should be done to understand the unique impact gaming’s intentional design has on these dynamics, and to explore lessons and models that may be transferable or avoidable.

As the community working to build a healthier digital world explores new models and ways of doing things, it should collaborate with and include the gaming community as a core part moving forward. Likewise, governments and civil-society leaders would be wise to pay closer attention to key parts of the gaming industry, as other governments seek to amass and exert power through gaming companies. These companies are spread throughout the world, building tools and content that will be used by billions of people for everything from expressing their views to running their businesses to innovating designs entirely for online spaces.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This annex reflects contributions from the following members of the Task Force for a Trustworthy Future Web: Rose Jackson, DFRLab; Kimberly Voll, Fair Play Alliance; Camille Francois, Niantic; and Sidney Olinyk, Duco, as well as the following contributing experts to the task force: Sara Grimes, University of Toronto, and Charlie Sinhaseni, Duco Experts. 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.

APPENDIX

GAMING ECOSYSTEM MATRIX

Due to differences in reporting practices and fiscal year calendars, the figures below are not directly comparable to each other. They are meant solely to provide a broad overview of the scale of revenue generated by different companies in the gaming ecosystem, and demonstrate the categories they have business in.

COMPANY	REVENUE (USD) 2022	HARDWARE	DEVELOPMENT	ENGINES	PUBLISHING	STOREFRONT	AR / VR / XR
TENCENT ¹	<u>32.2 BILLION</u>		▶		▶		▶
SONY	<u>18.2 BILLION</u>	▶	▶ PLAYSTATION STUDIOS		▶	▶	▶
APPLE	<u>15.3 BILLION</u>	▶			▶	▶	▶
MICROSOFT	<u>12.9 BILLION</u>	▶	▶ BETHESDA MOJANG		▶	▶	▶
GOOGLE	<u>11.08 BILLION</u>	▶			▶	▶	▶
NETEASE	<u>9.6 BILLION</u>		▶		▶		▶
ACTIVISION/BLIZZARD	<u>8.1 BILLION</u>		▶		▶	▶	▶
NINTENDO	<u>8.1 BILLION</u>	▶	▶		▶	▶	▶
ELECTRONIC ARTS	<u>6.5 BILLION</u>		▶		▶	▶	▶
EPIC GAMES	<u>6.23 BILLION</u>			▶	▶	▶	▶
SEA LTD. (GARENA)	<u>4.3 BILLION</u>		▶		▶	▶	▶
TAKE-TWO	<u>3.5 BILLION</u>		▶		▶		▶
SQUARE ENIX	<u>2.99 BILLION</u> ²		▶		▶		▶
ROBLOX	<u>2.2 BILLION</u>		▶		▶ SELF- PUBLISHED		▶
UBISOFT	<u>2.47 BILLION</u>		▶		▶	▶	▶
EMBRACER GROUP (THQ NORDIC)	<u>1.65 BILLION</u> ³		▶		▶	▶	▶
UNITY TECHNOLOGIES	<u>1.39 BILLION</u>			▶	▶		▶

¹ Tencent owns Funcom, Riot Games, and a number of other gaming properties. They also hold significant stakes in major companies like Epic.

² This figure is inferred; the company publishes "net sales" rather than revenue figures. Available stats do not reflect Embracer Group's recent acquisition of Square Enix-owned studios and IP.

³ Available figures do not reflect Embracer Group's recent acquisition of Square Enix-owned studios and IP.

ANNEX 5

SCALING TRUST ON THE WEB

COLLECTIVE SECURITY IN A FEDERATED WORLD

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB


ANNEX 5

COLLECTIVE SECURITY IN A FEDERATED WORLD

TABLE OF CONTENTS

Introduction	2
Moderating the Fediverse	3
The ABCs of Moderating Disinformation	6
Content	6
Behavior	7
Actors	9
Next Steps	11
Authorship and Acknowledgements	12

INTRODUCTION



Many discussions about social media governance and trust and safety—among regulators, developers, researchers, and users alike—are focused on a small number of centralized, corporate-owned platforms that currently dominate the social media landscape: Meta’s Facebook and Instagram, YouTube, Twitter, Reddit, and a handful of others. The emergence and growth in popularity of federated social media services, like Mastodon and Bluesky, introduces new opportunities, but also significant new risks and complications. While federated services continue to be dwarfed in size in comparison to platforms like Facebook and Twitter, the steady rise in their adoption warrants further attention and study. In the [case of Mastodon](#), for example, changes in ownership and governance at Twitter appear to have significantly accelerated the platform’s adoption, with some estimates showing more than ten million currently active users. For all the optimistic rhetoric that Mastodon is “like Twitter, but without the bad parts,” we should assume that centralized and decentralized platforms share a common set of threats from motivated malicious users—and require a common set of investments to ensure trustworthy, user-focused outcomes.

Broadly speaking, the “[fediverse](#)” is a catch-all term for a wide array of distinct products, services, and platforms that interconnect using a set of shared communication protocols such as the W3C standard [ActivityPub](#) or the under-development [Bluesky AT Protocol](#). In place of a centralized social media platform like Twitter, a federated alternative might involve dozens, hundreds, or even thousands of individual servers running instances of an open-source product. Despite being maintained by separate people or groups, servers using the same underlying protocol are interoperable, communicating with each other (and, in turn, allowing their users to access one another’s’ content). A number of distinct products have been built atop these decentralized standards, including Mastodon (a Twitter-like social media platform) and Pixelfed (an Instagram-like platform focused on media sharing).

These emergent distributed and federated social media platforms offer the promise of alternative governance structures that empower consumers and can help rebuild social media on a foundation of trust. Their decentralized nature enables users to act as hosts or moderators of their own instances, increasing user agency and ownership, and platform interoperability ensures users can engage freely with a wide array of product alternatives without having to sacrifice their content or networks. Unfortunately, they also have many of the same propensities for harmful misuse by malign actors as mainstream platforms like Face-

book and Twitter, while possessing few, if any, of the hard-won detection and moderation capabilities necessary to stop them. More troublingly, substantial technological, governance, and financial obstacles hinder efforts to develop these necessary functions.

This paper offers an assessment of the trust and safety (T&S) capabilities of federated platforms—with a particular focus on their ability to address collective security risks like coordinated manipulation and disinformation.¹ We focus on disinformation risks for two reasons. First, they have significant societal impact. Second, disinformation threats primarily are detectable and mitigable as actor- and behavior-level phenomena, rather than the content-level moderation approaches discussed in most research about trust and safety.

Beginning with a broad review of the current structures and practices of moderation on federated services, we examine the particular issues created by persistent, adversarial campaigns. We identify several significant structural impediments to robust mitigation of disinformation threats, given current technical and labor models of moderation: namely, the shortcomings of content-driven approaches to moderation in counteracting these campaigns, and the obstacles to implementing behavioral defenses.

MODERATING THE FEDIVERSE

Most discussions of fediverse moderation have, reasonably, focused on the essential contrast between centralized, corporate approaches to content governance (like those employed by Meta, Google, and Twitter), and a distributed, community-driven approach native to federated services like Mastodon. The essential feature of federated systems, and of the protocols like ActivityPub underlying them, is decentralization. Each instance of a federated service can choose for itself what its governance approach will be; in turn, its governance decisions extend only so far as the (virtual) boundaries of that particular server. As Alan Rozenshtein [summarizes](#), “No instance can control the behavior of any other instance, and there is no central authority that can decide which instances are valid or that can ban a user or a piece of content from the ActivityPub network entirely. As long as someone is willing to host an instance and allow certain content on that instance, it exists on the ActivityPub network.” By design, the perimeter of the fediverse is highly permeable; new platforms and users can enter and exit federated systems readily, to both the benefit and detriment of the overall network.

Despite a lack of protocol-mandated governance, many of the more populous parts of the fediverse engage in at least some form of moderation. For example, the Mastodon Server Covenant (which governs whether a Mastodon instance is listed in the central server picker maintained by Mastodon’s creator) requires “active moderation against racism, sexism, homophobia and transphobia.” While a comprehensive assessment of the policies of federated platforms (including the legitimacy of those policies, and their sufficiency in protecting speech and user safety) is beyond the scope of this article, it is worth noting that where they do exist, the community standards of fediverse instances are often sparse, high-level statements of principle, rather

¹ The terminology used to describe campaigns like the Russian Internet Research Agency (IRA) targeting the 2016 US elections is complex, and increasingly politicized—with terms like “disinformation” now broadly associated with allegations of ideological censorship by technology platforms. Broadly, we use the terms “disinformation,” “information operation,” “platform manipulation,” and “coordinated manipulation” interchangeably throughout this article—though they each refer to slightly different phenomena. More specifically, we draw on a taxonomy of the forms of information disorder originally developed by First Draft, which defines “disinformation” as “content that is intentionally false and designed to cause harm” and “malinformation” as “genuine information that is shared with an intent to cause harm.” As we discuss in this article, specific adversarial campaigns like the IRA’s efforts may involve a mixture of deceptive behaviors, outright lies, and true information shared to mislead or polarize. In part because of these ambiguities, technology platforms have developed alternate terms—such as Facebook’s “coordinated inauthentic behavior”—that characterize such campaigns by the use of deceptive practices like operating multiple social media accounts. As researcher Evelyn Douek has noted, these terms can also be problematic, in large part because of how platform specific they can be, and the difficulties of auditing the standards used by platforms to implement them.

FEDERATED PLATFORMS STATE ASSESSMENT

	CENTRALIZED	FACEBOOK	INSTAGRAM	HORIZON WORLDS	TWITTER	REDDIT	YOUTUBE	DECENTRALIZED	MASTODON	PIXELFED	DIASPORA	PEERTUBE
POLICY												
Public community norms / standards (high level statements)		▶	▶	▶	▶	▶	▶		▶	▶	▶	▶
Public policy explanations with enforcement criteria (detailed)		▶	▶	×	▶	×	▶		×	▷	×	▷
Behavioral manipulation / CIB / platform manipulation policy		▶	▶	×	▶	▷	▶		×	×	×	×
REPORTING												
User reporting capabilities for policy violations		▶	▶	▶	▶	▶	▶		▶	▶	▷	▶
ENFORCEMENT CAPABILITIES												
Permanent account bans		▶	▶	○	▶	▶	▶		▶	▶	▶	▶
Temporary account bans / timeouts		▶	▶	○	▶	▶	▶		▶	▶	×	▶
Ban evasion detection		▷	▷	○	▷	▷	▷		▷	▷	×	▷
Post/content deletion		▶	▶	-	▶	▶	▶		▶	▶	▶	▶
Account visibility restriction		▶	▶	-	▶	▶	▶		▶	×	×	▶
Post/content visibility restriction		▶	▶	-	▶	▶	▶		▶	▶	×	▶
Demonetization		▶	▶	▶	▶	-	▶		-	-	-	-
Automated enforcement tools (heuristics, ML)		▶	▶	×	▶	▶	▶		×	×	×	×
URL blocking		▶	▶	-	▶	▶	▶		×	×	×	×
Media hashing/matching		▶	▶	-	○	▶	▶		×	×	×	×
User-facing moderation controls (block, mute, etc)		▶	▶	▶	▶	▶	▶		▶	▶	×	▶
User identity verification (ID checks, etc)		▶	▶	▶	▶	×	▶		×	×	×	×
Antispam challenges (reCAPTCHA, phone verification)		○	○	-	▶	×	▶		▷	▷	▷	▷
Defederation / instance blocking		-	-	-	-	-	-		▶	○	×	▶
TRANSPARENCY												
Published transparency report		▶	▶	×	▶	▶	▶		×	×	×	×
Terms of service enforcement data		▶	▶	×	▶	▶	▶		×	×	×	×
Behavioral manipulation / CIB / platform manipulation data		▶	▶	×	▷	▶	×		×	×	×	×
Legal information requests data		▶	▶	×	▷	▶	▶		×	×	×	×
Legal removal demands data		▶	▶	×	▷	▶	▶		×	×	×	×
Country/jurisdictional breakdowns of data		▶	▶	×	▷	▶	▶		×	×	×	×
APIs												
Publicly available GET APIs for core platform data (posts, users)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶
Publicly available POST APIs for core platform functions (posts, etc)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶
Publicly available moderation APIs (block, mute, etc)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶

RATING	DESCRIPTION
▶ COMPLETE	The existence of the capability is publicly documented, and is available for use (or has demonstrably/documentably been used) at scale across the platform's core products/business units. Click on icon for hyperlinked citation.
▷ PARTIAL	The capability exists, but (1) is not applicable to all of the platform's core products/business units, or (2) has significant functionality gaps that prevent effective use for moderation. Click on icon for hyperlinked citation.
×	None
-	N/A
○ LIKELY EXISTS	The platform likely possesses the capability, but does not have publicly-listed information on it. Click on icon for hyperlinked citation.

than the detailed policies published by larger, centralized platforms. This creates practical ambiguities for the people responsible for moderating content, as well as uncertainty for users about precisely what goes in a given context.

To implement these policies, most federated platforms provide instance administrators and moderators with a rudimentary set of moderation tools. Mastodon, for example, allows moderators to ban individual users, as well as to delete or restrict the visibility of individual pieces of content and accounts.

Federated moderation differs from centralized moderation in an essential way by virtue of the distribution of accounts across multiple instances. A user account has a “local” instance on which it resides, and that instance’s moderators have the ability to take direct, destructive action on the user’s content (such as deleting it). But, for non-local accounts and content—that is to say, users whose accounts reside on other instances—administrators can only impact their local copies of that content, which influences only the experiences of local users. If a user on instance A encounters a harmful post from a user on instance B, instance A’s moderators have no ability to compel instance B to take any action on the harmful post. Mastodon’s tools do, however, allow local moderators to take instance-level action to restrict the visibility of content and accounts for all of their local users of an instance, even if that content is permitted on other instances. This has the beneficial effect of giving users greater choice about the policies and governance approaches influencing what they see on social media—but it also makes it more challenging to address fediverse-wide risks created by instances that either cannot or choose not to moderate, whose harmful effects may persist through online and offline action by instance users, even if they are cordoned off from other parts of the fediverse through technical blocks.

In order to address the risks created by specific instances that fail to moderate appropriately, most federated services offer administrators the ability to take moderation action at the instance level, impacting all users on a remote instance, instead of just moderating post by post or account by account. In the case of Mastodon, for example, instances are able to defederate themselves from other servers—in essence, refusing to communicate with or display content from a server deemed to be problematic, rendering its content invisible for all the users on an instance that has chosen to defederate from it. Defederation is largely a server-by-server decision, and, beyond a small number of shared blocklists, few technical or community capabilities exist to deploy these measures at scale across tens of thousands of separate federated instances. Nevertheless, these tactics have, at times, been deployed as a form of broad-based, collective action by fediverse instance moderators—including the notable [case](#) of broad defederation from extremist platform Gab.

While defederation offers one scaled mechanism for addressing repeated or prolific harmful conduct, federated platforms largely lack industry-standard capabilities for broad or automated content moderation. Mastodon, for example, does not provide moderators with the ability to block harmful links from being shared on the service. This prevents moderators from being able to ingest lists of known-bad URLs (such as spam and phishing sites) in order to programmatically restrict them. Mastodon also lacks essential tools for addressing media-based harms, like child sexual exploitation, such as media hashing and matching functions (although a number of third parties, including [Cloudflare](#), make such tools available to customers of their content-delivery services). Critically, many of the existing federated platforms have not implemented moderator-facing tools for deploying automation and machine learning to streamline and scale repeated content-moderation actions—functions that are an essential part of the moderation toolkit at all of the existing large, centralized platforms.

As a consequence of the nascent state of moderation capabilities—in terms of both technical features and the manual practices of implementing them—clear governance challenges have emerged for federated service admins. Making moderation a distributed challenge means each instance operator has to reinvent many of the policies and procedures of moderation for themselves. As Ben Werdmuller [puts it](#), “While software is provided to technically moderate, there are very few ecosystem resources to explain how to approach this

from a human perspective.” The results are predictable for anyone familiar with the challenges of social media content moderation. Users [report](#) erroneous or inexplicable bans, with limited recourse from volunteer admins moonlighting as content moderators. Larger-scale harassment [campaigns](#) can overwhelm victims and admins alike. Driven by business imperatives, virtually all centralized platforms at least attempt to mitigate these harmful behaviors. But, absent the financial support that goes along with centralized, corporate social media, few parts of the fediverse have been able to successfully marshal the human and technological resources required to successfully execute proactive, accurate content moderation at scale.

THE ABCS OF MODERATING DISINFORMATION

The existing difficulties of moderating federated systems—chief among them, a general lack of resourcing supporting these efforts, even as federated platforms like Mastodon continue to see growth in their adoption—are exacerbated by the highly adaptive nature of coordinated manipulation threats, and the fact that they often require quite different approaches to moderation than those for abuse or hate speech.

To understand these challenges, it’s helpful to break down the problem of moderating disinformation into a few components, which researcher Camille François helpfully [taxonomized](#) as the “ABCs”: actors, behaviors, and content.

CONTENT

Nearly all content-moderation discussions begin with the “C” in François’s ABC framework: the content being shared. These analyses focus on the content of a post or account: the language it uses, the links it shares, and the characteristics of a profile. The fundamental challenge of disinformation, however, is that it’s seldom apparent from content alone that what you’re looking at is actually part of a manipulative campaign. The post- and profile-level evidence most directly available to moderators is rarely dispositive.

Looking back at key [examples](#) of the Russian Internet Research Agency (IRA) activity on Twitter in 2016, what is most striking about posts from prominent accounts like “Crystal Johnson,” a Russian persona purporting to be an African-American woman, is that, by and large, the content of the posts was true: while the IRA’s earliest efforts involved comically ineffective rumormongering about an alleged outbreak of Ebola in Atlanta, the bulk of its [activity](#) during and after the 2016 US elections used a more subtle tactic of sharing factually accurate but divisive rhetoric using inauthentic behaviors (such as fake accounts). This stymied efforts to moderate content based on policies that evaluate the substance of a post.

Even when we know content has been created by a troll farm, addressing it as content is challenging (if not impossible). For example, researcher Josh Russell [captured](#) hundreds of examples of memes created by the IRA on Instagram in 2018; those same exact memes [resurfaced](#) a year later in a network of spammy Facebook pages operated out of Ukraine. If Meta, possessing all the relevant data about these campaigns and having extensive capabilities to detect similar media, couldn’t catch this, how can we expect Mastodon instance moderators to keep pace, particularly given the lack of media-hashing and matching functions? And even if these capabilities were developed and implemented, matching-based approaches are inherently reactive to already-known examples of disinformation; modifications to existing assets, or the creation of novel content, would quickly undermine the effectiveness of these approaches.

These challenges are exacerbated by the phenomenon of real people authentically sharing content from, or similar to, trolls. Many of the memes originally created by IRA staff in St. Petersburg continue to circulate on social media among folks who just happen to think they’re funny or interesting. If real people intentionally choose to amplify messaging sourced from, or consistent with, a government-sponsored trolling campaign, it’s not obvious what, if anything, moderators should do. While some have argued that stronger media liter-

acy would greatly help with this issue, the particular characteristics of IRA-style inauthenticity makes this a challenging proposition. What forms of literacy would help most people recognize a Crystal Johnson-style account as inauthentic using publicly available data? The sparseness of profiles on services like Mastodon, as well as their relative newness, makes it harder to make genuinely informed judgments—limitations that apply to users and moderators alike.

BEHAVIOR

At the core of disinformation campaigns is the concept of manipulative behavior: the practice of engaging in tactics of sharing and disseminating content that seek to inauthentically propagate, promote, or inflate the reach of an account or piece of content. As François puts it in the ABC framework: “At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have perceived impact that a greater number of actors would have if the campaign were organic.” Put another way, a few staffers at a troll farm in St. Petersburg are unlikely to be particularly influential absent behavior that skews the attention economy of social media in their favor.

In many cases, this is just another way of referring to spam.² High-volume, low-sophistication political-manipulation campaigns became a feature of Twitter in particular—with threat actors based in [Venezuela](#) and [China](#) (to give just two examples) deploying them with some regularity.

Federated services, at least in their present implementations, have some inherent resilience to these tactics. The lack of algorithmic recommendations means there’s less of an attack surface for inauthentic engagement and behavioral manipulation. While Mastodon has introduced a version of a “trending topics” list—the true battlefield of Twitter manipulation campaigns, where individual posts and behaviors are aggregated into a prominent, platform-wide driver of attention—such features tend to rely on aggregation of local (rather than global or federated) activity, which removes much of the incentive for engaging in large-scale spam. There’s not really a point to trying to juice the metrics on a Mastodon post or spam a hashtag, because there’s no algorithmic reward of attention for doing so. The lack of built-in monetization programs on virtually all federated platforms—at least presently—likewise reduces incentives for programmatic malfeasance.

These disincentives for manipulation have their limits, though. Some of the most successful disinformation campaigns on social media, like the IRA’s use of fake accounts, relied less on spam and more on the careful curation of individual “high-value” accounts—with uptake of their content being driven by organic sharing, rather than algorithmic amplification. Disinformation is just as much a community problem as it is a technological one (i.e., people share content they’re interested in or get emotionally activated by, which sometimes originates from troll farms)—which can’t be mitigated just by eliminating the algorithmic drivers of virality.

Detection of behavioral manipulation relies, in large part, on access to data about on-platform activity—and the openness of federated platforms has largely resulted in the ready availability of application programming interfaces (APIs) to enable this kind of access. For [example](#), Mastodon has a robust set of public APIs that would allow researchers to study the conversations happening on the service. But federation complicates the use of these APIs to study ecosystem-level threats. Whereas Twitter’s APIs offer a single channel for col-

² In the early days of Congress investigating Russian interference in the 2016 US election, Twitter staff briefed stakeholders on Capitol Hill about the company’s efforts to combat what we were calling “political spam.” We were excoriated by a few of the people with whom we spoke, who said that even calling it “spam” meant we were missing the gravity of the situation. Twitter subsequently came up with the term “platform manipulation” as an alternative that would signal how seriously we took the issue. See: Patrick Conlon, William Nuland, and Kanishk Karan, “Investigating Influence Operations By Twitter Integrity,” in Victoria Smith, Jon Bateman, and Dean Jackson, eds., [Perspectives for Influence Operations Investigators](#) (Washington DC: Carnegie Endowment for International Peace, 2022).

lecting data about all the activity happening globally across the Twitter service, Mastodon's APIs are mostly instance specific. As a result, many data-collection efforts either involve focusing on a handful of the largest instances, or needing to go down an essentially limitless rabbit hole of collecting data from successively smaller and smaller instances until you reach a point of diminishing returns—with no guarantee that the threats you're hunting aren't lurking on the $n+1$ th instance from which you'd collect data.

The federated nature of this threat creates similar challenges for moderators. Many of the techniques employed by large platforms to detect manipulation involve surveying the full population of accounts and activity, and looking for unusual clusters or patterns of behavior within that population—a practice of threat identification using centralized telemetry. To give a rudimentary example: if you look at posts containing a hashtag like #ElectionNight2022, group those posts by the Internet Protocol (IP) address from which they were sent, and observe that a bunch of them were sent from an IP address in Russia, you might investigate the accounts responsible to see if something fishy is going on. But in a federated system, instance admins only have comprehensive logs for the activity of local users of their particular instance—which means a threat actor who spreads their inauthentic accounts across a handful of the biggest instances is both less likely to be caught as behaviorally anomalous and less likely to have the full scope of their operation, across all the instances on which they operate, be detected. An analyst is less likely to spot a suspicious cluster of accounts if it's just one or two users among tens of thousands.

This also assumes that instance moderators have the time, knowledge, tools, and governance frameworks necessary to do the highly specialized work of disinformation detection and analysis. Training programs at large platforms to get even technically proficient analysts fully up to speed on advanced analytic techniques can take months. There are also costs beyond just time and attention. Even if you assume that moderators have the necessary technical skills to do this work, the compute costs alone of querying against these large-scale datasets are considerable; we can't reasonably expect volunteers to do this work pro bono. These capabilities beget fraught privacy and user-control challenges as well. What safeguards exist to ensure instance admins, or their designees, engage only in appropriate uses of sensitive user logs?

Finally, there's also the challenge of how, exactly, to moderate a distributed but coordinated threat. Mastodon's moderation capabilities provide for a few rudimentary anti-spam techniques for addressing scaled threats (techniques even the Mastodon documentation notes will be circumvented by dedicated spammers)—but Mastodon moderation is focused largely on either dealing with individually problematic users (by restricting or banning them from a given instance) or the radical option of defederating a wholly problematic instance. Spam and platform manipulation are unlikely to be solvable using this tactic, because they primarily manifest as distributed threats across mainstream, non-malicious instances. Put another way, we shouldn't expect sophisticated adversarial threats to concentrate themselves on single instances, waiting to be defederated. Instead, inauthentic accounts are likely to be dispersed across mainstream servers. This creates a distributed burden of detection across already-overworked and under-equipped moderators, who need to deal with these accounts one by one, instance by instance.

This is a social challenge as well as a technical one. The ActivityPub protocol gives instance operators ways to defend themselves against bad actors by defederating problematic servers, but what is the appropriate course of action when well-intentioned admins may be unaware of, or unable to meaningfully address, the malicious activity they host? As with the challenges of addressing the attempted infiltration of social movements by troll farms, it isn't always clear how to know which instances and individuals are trustworthy—a dynamic that malign actors can exploit, as happened with a Ukrainian Mastodon instance following the Russian invasion of Ukraine. At their worst, these dynamics may lead to otherwise legitimate instances being defederated or restricted for failing to appropriately moderate—to the detriment of their other legitimate users.

Among the most commonly proposed solutions to these issues is making Mastodon instances invite only, or requiring some kind of trusted referral model for new signups. This may well be a viable solution for parts of

the fediverse that intentionally prioritize small community size and affinity based on identity or interest. But the “gated community” model has at least three key challenges as a broader strategy. First, this only solves the problem of “local” manipulation, not the impacts of federated behavior on non-local viewers of that content. Second, it’s not clear that this is actually a way to address the most sophisticated and insidious forms of manipulative behavior. Elaborately constructed inauthentic profiles—like Crystal Johnson, or the deep, cross-platform persona development tactics [described](#) in a recent expose about an Israeli disinformation purveyor—will often withstand anything but the most invasive forms of validation. (And, inevitably, the more invasive validation becomes, the less usable a service is by vulnerable people and groups, who might have good reasons for not wanting to disclose their personal information to instance operators they don’t know or trust.) Finally, and most fundamentally, for people looking to Mastodon and the fediverse [as an alternative](#) to centralized social platforms like Twitter, raising barriers to entry introduces fundamental tradeoffs against the very network effects that could help make Mastodon a mass-market product.

ACTORS

In François’s framework, “actors” refers to the people or groups behind deceptive activity. The basic premise is that it matters who or what is engaged in malicious or harmful conduct. Often, actor-level analysis is reduced to the security practice of [“attribution”](#)—the name-and-shame exposure of the individuals or groups responsible for an attack or intrusion. Certainly, attribution has an important part to play in counteracting disinformation; it gives nation-states critical evidence needed to enact offline consequences for online malfeasance. We don’t need to look further than the dozens of indictments of Russian operatives following the 2016 elections to see this direct interplay of platform-based investigations and offline law-enforcement action.

But attribution is far from the only goal of actor-level analysis. Understanding the actors responsible for a disinformation campaign meaningfully influences how platforms respond—and can give platforms necessary tools for addressing these challenges in a scalable way. The present state of fediverse moderation—from both labor and technological perspectives—has two primary structural constraints on actor-level analysis: a lack of capability and capacity for longitudinal enforcement; and a lack of collaboration with other groups tracking the same threat actors, which results in inefficiency and detection gaps.

Longitudinal analysis refers to tracking and analyzing the behavior of specific threat actors or patterns of malicious behavior over time. These practices have typically been carried out by platforms themselves, and by a wide array of civil-society and academic groups. These are not abstract pursuits; they have specific, practical applications that meaningfully contribute to platform capabilities to mitigate disinformation. Principally, understanding the behaviors and motivations of persistent threats helps platforms develop effective mitigation strategies suited to applying optimal, cost-effective pressure to a particular actor based on their unique goals and constraints. For example, cost-optimizing threat actors—like many of the commercially motivated groups peddling political spam in order to sell t-shirts or redirect traffic to ad-filled content farms—will be most impacted by enforcement strategies that raise the cost of doing business. The individual unit costs of spam are low, but so, too, are the gains realized through these campaigns. Strategically imposing additional expenses through mechanisms like mandatory phone-number verification, completion of CAPTCHA challenges, and domain blocking can, over time, make it cost-ineffective for financial spam campaigns to target an effective platform. But these enforcement measures have user-experience and privacy tradeoffs, and platforms generally avoid applying them indiscriminately for fear of alienating users—which requires a targeted approach that seeks out and enforces against specific threats in specific ways.

The current state of fediverse moderation has two key constraints on the ability to enact this kind of targeted pressure on adversarial behavior. First, actor-level analysis requires time-consuming and labor-intensive tracking and documentation. Differentiating between a commercially motivated spammer and a state-backed troll farm often requires extensive research, extending far beyond activity on one platform or website. The

already unsustainable economics of fediverse moderation seem unlikely to be able to accommodate this kind of specialized investigation.

Second, even if you assume moderators can, and do, find accounts engaged in this type of manipulation—and understand their actions and motivations with sufficient granularity to target their activity—the burden of continually monitoring them is overwhelming. Perhaps more than anything else, disinformation campaigns demonstrate the “persistent” in “advanced persistent threat”: a single disinformation campaign, like Chi-na-based Spamouflage Dragon, can be responsible for tens or even hundreds of thousands of fake accounts per month, flooding the zone with low-quality content. The moderation tools built into platforms like Mastodon do not offer appropriate targeting mechanisms or remediations to moderators that could help them keep pace with this volume of activity. Moderation actions are wholly manual, and are limited to either banning or restricting individual accounts, or blocking entire ranges of IP addresses or email domains. Moderators lack the capability to deploy heuristics—essentially, sets of rules that describe patterns of adversarial behavior—that can automate these actions. Without these capabilities to automate enforcement based on long-term adversarial understanding, the unit economics of manipulation are skewed firmly in favor of bad actors, not defenders.

These are solvable product and engineering challenges—and no doubt the moderation tools built into Mastodon and other federated products will improve over time. But there are critical labor components as well. In the case of a persistent but poorly obfuscated campaign like Spamouflage Dragon, detection isn’t especially difficult. But as the tactics and thematic focus of the campaign evolves over time, scaled remediation requires continual maintenance to keep things from going off the rails. Heuristics that are viable one day can become inaccurate the next. Machine-learning models exhibit drift over time and can either under-detect or over-detect the target activity. “Set it and forget it” is not a viable strategy for dealing with dedicated adversaries. Look no further than Twitter following the dismissal of staff responsible for [monitoring](#) Chinese-language disinformation, heuristics developed to address Spamouflage Dragon and other campaigns have, according to reports, [declined](#) in accuracy to a point where the legitimate accounts of activists and users are being inaccurately restricted or banned. Responsible deployment of even sophisticated technical enforcement capabilities requires ongoing, sustained effort by moderators.

Critically, efforts to disrupt persistent threat actors are most successful when approached at a community or ecosystem level, rather than by individual platforms in isolation. Many of the most prolific disinformation campaigns are notable for their [presence](#) across multiple platforms. Russian campaigns targeting the White Helmets in Syria in 2017 and 2018 spanned Twitter, YouTube, and mainstream and alternative media properties. More recently, the Secondary Infektion operations [promoting](#) Russian interests spanned more than three hundred sites and platforms. From an analytic perspective, it can be challenging, if not impossible, to recognize individual accounts or posts as connected to a disinformation campaign in the absence of cross-platform awareness of related conduct. The largest platforms—chiefly, Meta, Google, and Twitter (pre-acquisition)—regularly shared information, including specific indicators of compromise tied to particular campaigns, with other companies in the ecosystem in furtherance of collective security. Information sharing among platform teams represents a critical way to build this awareness—and to take advantage of gaps in adversaries’ operational security to detect additional deceptive accounts and campaigns.

Federated moderation makes this kind of cross-platform collaboration difficult. Thousands of individual instance operators each have responsibility for a potential target of this conduct, but it’s infeasible for larger platforms, like Meta and Google, to engage with moderators or admins from each instance directly. Even assuming these engagements are limited to a handful of the largest fediverse instances, the legal frameworks and contractual protections needed to share data across platforms without running afoul of privacy regulations like the General Data Protection Regulation (GDPR) require specialized legal expertise and negotiation, which are often out of reach for hobbyist efforts. In addition, absent an institutionalized way to verify the trustworthiness and legitimacy of instance admins and moderators, larger platforms will have limited information

on who they are working with—and, correspondingly, may either choose not to engage or feel constrained in their ability to share relevant data. Bad actors posing as moderators of legitimate fediverse instances can leverage these structural ambiguities to gain access to larger platforms’ staff and intel, creating commercial, political, and privacy risk.

Even among federated services, it’s challenging for instance moderators to engage with each other in a structured way to counteract shared threats. Collaborative security models are common both within the social media industry and outside of it—including [financial-intelligence units](#) in the financial-services sector, and [information-sharing and analysis centers](#) in the information security context. These institutionalized collaborations are predicated on a high degree of alignment about the scope and nature of the threats in question. While decentralized community governance has had notable successes on platforms like Wikipedia, the notable lack of agreement on norms and standards across instances makes it challenging for these collaborative practices to adapt to the fediverse. For example, on Mastodon, [tagging](#) discussions using “#fediblock” has emerged as a grassroots practice for sharing information about bad actors, but these approaches have run up against the challenges of a fully distributed, low-trust model. Moderators report that it’s hard to know which accounts of bad behavior are trustworthy or verified enough to warrant enforcement without firsthand confirmation.

This is a product not only of technical capabilities, but of the cultural norms of federated systems, exposing a core challenge in establishing effective collaboration. Evelyn Douek has [written](#) critically about the so-called “content cartels” that form when mainstream platforms collaborate with each other; federated approaches, in part, offer an alternative configuration. When platforms are designed and built to empower individuals and communities to be self-sovereign, as in the case of many federated services, their operators and moderators may be reasonably skeptical of the kind of centralized designations inherent in this kind of information sharing. It would hardly be desirable for instance operators to uncritically import enforcement decisions and bad-actor designations without ensuring that these data are both aligned with instance policies, and are sourced from trustworthy moderators. A failure state for the promise of the fediverse is homogeneity of moderation as a product of convenience. But leaving it to individual moderators to assess, designate, and track troll farms and other bad actors for themselves is hardly a reasonable alternative.

Establishing mechanisms for transparency reporting in the fediverse could help address difficulties in moderating across instances. This may soon become necessary, as the European Digital Services Act is likely to [classify](#) Mastodon instances as independent “online platforms” subject to transparency reporting obligations. No such reporting practice currently exists on major fediverse platforms, and the creation thereof would only be complicated by the need for compliance and coordination across moderators and admins, and the lack of a centralized structure to report on this information.

NEXT STEPS

As consumers explore alternatives to mainstream social media platforms, malign actors will migrate along with them—a form of cross-platform regulatory arbitrage that seeks to find and exploit weak links in our collective information ecosystem. Further research and capability building are necessary to avoid the further proliferation of these threats.

A critical element of this work is identifying the specific intervention strategies that are suited to the sociotechnical properties of federated and distributed social platforms. Key research questions include the following.

- ▶ What are the risks and challenges posed by disinformation and manipulative behavior on federated platforms, and how do these risks differ from those created on centralized social media services?

- ▶ What policies and governance approaches currently exist for manipulative behavior and disinformation across major instances of federated services?
- ▶ How do the users and operators of federated services conceptualize the risks of manipulative conduct, given the norms and governance structures of existing decentralized communities?
- ▶ What are the existing moderation capabilities built into federated services, and how effective are they at addressing behavioral and scaled threats?
- ▶ What technical capabilities—moderation tools, datasets, APIs, etc.—are required to effectively manage coordinated manipulation and disinformation threats?
- ▶ What moderation and analytic capabilities are necessary to help instance operators, moderators, and the users of federated services address the risks and threats created by persistent adversarial behavior?
- ▶ What are the appropriate governance frameworks and organizational structures for this work in a decentralized context?

Answers to these questions will help structure responses across three critical constituencies: the developers of open-source fediverse services, and the developers of complementary tools and features that enable effective moderation of federated social media; the individuals and groups engaged in investigations, analysis, and moderation of federated services; and investors, funders, and donors engaged with platform governance and counter-manipulation efforts.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This report was authored by Yoel Roth and Samantha Lai. Yoel Roth is a technology policy fellow at the University of California, Berkeley, and a nonresident scholar at the Carnegie Endowment for International Peace. He was previously the head of trust and safety at Twitter, and received his PhD from the Annenberg School for Communication at the University of Pennsylvania. Samantha Lai is a research analyst with the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace. Prior to joining Carnegie, she was a research analyst at the Brookings Institution and the Foreign Policy Research Institute. This work was supported by the William and Flora Hewlett Foundation, the John S. and James L. Knight Foundation, and the Carnegie Endowment for International Peace. The authors are grateful to Patrick Conlon, Renee DiResta, Camille François, Jeff Jarvis, Jaz-Michael King, Hilary Ross, and members of the Atlantic Council Task Force for a Trustworthy Future Web for their feedback and perspectives.

ANNEX 6

SCALING TRUST ON THE WEB

LEARNING FROM CYBERSECURITY, PREPARING FOR GENERATIVE AI

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB


ANNEX 6

**LEARNING FROM CYBERSECURITY,
PREPARING FOR GENERATIVE AI**

TABLE OF CONTENTS

Introduction	2
Lessons Learned from Cybersecurity’s Evolution	3
Education, Professional Training, and Research	3
Narrative and Storytelling	4
Information Sharing, Identifying Threats, and Measuring Harms	5
Scrutiny, Politicization, and Risk Tolerance	7
Decision-making, Leadership, and Success	9
Counterbalancing Global North Dominance	9
Looking Ahead to Generative AI	10
Generative Technologies and the Industry Outlook for Trust and Safety	10
Generative AI: Friend or Foe to Content Moderation?	10
Conclusion	12
Authorship & Acknowledgments	12

INTRODUCTION



As the nascent trust and safety (T&S) field develops, it is uniquely positioned to develop with an intentional focus on leveraging lessons that have been learned through the development of adjacent fields, such as cybersecurity. The formalization of the field also allows for more coherent forecasting and prioritization, as emerging technologies like generative artificial intelligence (GAI) create opportunities for extreme risks, and also potential new solutions to longstanding T&S challenges.

Cybersecurity is a relatively young field that has rapidly matured over the past two decades. Whereas twenty years ago, few nonexperts knew what a hack or breach was, cybersecurity is now front-page news around the world. While thorny policy problems—e.g., the [encryption debate](#)—persist, what was once largely an insular technical field has evolved into a multidisciplinary and multisectoral ecosystem.

Cybersecurity has much to offer the younger T&S field, in large part due to the maturity gap between the two communities. Like any rapidly maturing field, cybersecurity has both successes to emulate and failures to avoid repeating. Across dimensions like education, professionalization, risk management, and vendor capacity, cybersecurity has developed pathways that could accelerate the development of the T&S field, if emulated. By the same token, some consistent failings within cybersecurity—especially with respect to [diversity, equity, and inclusion](#)—can serve as a cautionary tale and incentivize different approaches as T&S matures.

Meanwhile, policy, practice, business models, and threat models for GAI have been changing by the day since ChatGPT was publicly released in November 2022. While it is not clear how this technology or its use will evolve, it is clear that its impact will be transformational. As a range of GAI tools are being unleashed for widespread public and commercial use, it is both possible and important to forecast ways in which this technology could be leveraged—positively and negatively—within T&S.

This annex seeks to illuminate where T&S can learn from cybersecurity, while still charting a nuanced path based on the unique needs and circumstances inherent to the growing T&S field. The cybersecurity examples given are not exhaustive. Rather, they serve to highlight promising areas of inquiry for future research, design of new institutions, and overall field building. In addition, this annex provides a brief but specific examination of how GAI could influence content moderation practices, with the aim of showing the value of forecasting for broader T&S implications, and illuminating its impact on one of the most consistently challenging areas of T&S practice.

LESSONS LEARNED FROM CYBERSECURITY'S EVOLUTION

EDUCATION, PROFESSIONAL TRAINING, AND RESEARCH

Cybersecurity has made meaningful strides in the past decade but is not a monolithic field. Rather, it comprises a diverse array of communities, stakeholders, and practitioners with different backgrounds and perspectives that enjoy different levels of maturity in different areas. It is a useful comparison point for T&S given both fields' need to balance technical and social disciplines while serving the needs of business and society alike.

The past decade has seen cybersecurity develop a more robust workforce pipeline, with educational programs (e.g., specialized university and associate degrees, etc.) as well as a dizzying array of professional [certifications](#). Educational programs range from purely technical programs to [multidisciplinary/policy-oriented](#) ones. In addition, the US government developed the Workforce Framework for Cybersecurity (the [NICE Framework](#)) to help employers develop their cybersecurity workforce by establishing a common lexicon for cyber roles across sectors.

Creative, [team-based and immersive learning programs](#) have also taken root. The talent pipeline has been elongated to draw younger and more diverse individuals from earlier grades (especially high school) into the field through [age-appropriate](#) programming and mentoring. For example, [cyber.org](#) develops and offers free curricula and modules for K-12 teachers to use for teaching their students about cybersecurity. Some of these modules are also interdisciplinary, teaching students about cybersecurity and digital citizenship.

The rich academic ecosystem in cybersecurity extends beyond education to research. Researchers have long convened at a range of leading conferences (e.g., [USENIX](#), etc.) and have published in various journals (e.g., various [Association for Computing Machinery](#) journals and Institute of Electrical and Electronics Engineers symposia on security and privacy, etc.). These outlets have further expanded cybersecurity-specific branches and subprograms over the past two decades, such as the creation of the Workshop On Offensive Technology ([WOOT](#)). Moreover, the research community involves distinct subcommunities that partially overlap: academic researchers, private sector researchers, and security researchers (who oftentimes self-identify as hackers). Vendors and ethical hackers play a critical role in pushing for transparency and best practices overall. At conferences that cater to the hacker community, most notably [DEF CON](#), collaborative [hackathons](#) to solve technical and/or cyber policy problems are routine. More private sector-oriented conferences, such as [Black Hat](#) and [RSA](#), have less of a research component, but serve as critical venues for vendors and customers to meet and transact business.

There are several promising features of the cybersecurity field that could be emulated to aid the maturation of the T&S field. In support of developing a stronger pipeline, a NICE-like framework that articulates the full range of T&S roles, skills, and competencies across all sectors of society (beyond just companies to include regulators, civil society, etc.) could support workforce development, recruitment, and related talent-building efforts. Clearer parameters and components for focused T&S educational programs for high school, community college, university, and graduate-level programs should be defined. For example, [Stanford University](#) has launched a handful of T&S-focused courses and is coordinating a consortium of other interested schools, with shared educational resources, to deepen T&S studies; this is a promising effort that will hopefully expand to other universities globally, over time. Professional certifications for various T&S-focused skills (e.g., data science, content moderation, etc.) and knowledge areas (e.g., bullying and harassment, child sexual abuse materials, etc.) will also be important to develop. The field would benefit from a T&S-focused organization stepping up and taking the lead on certifications, much as the [SANS Institute](#) did for cybersecurity.

T&S leaders do warn that increasing academic requirements for T&S professionals could cut against some of the great strengths of T&S; it will be important to strike a balance so that T&S does not become the domain of the elite. Front-line content moderators, for example, may not come from university backgrounds,

but bring important knowledge and expertise to the field. In addition, the interdisciplinary backgrounds feeding thought into current T&S teams are widely seen as a great strength of the field, and necessary to its successful maturation.

T&S also has room to grow in terms of research conferences and journals. The [Trust and Safety Research Conference](#) is off to a promising start, as is [TrustCon](#). There is substantial room to grow before contending with the massive size and challenges of events drawing 30,000 to 50,000 people like DEF CON or RSA. T&S hackathons and other collaborative efforts to solve T&S problems and share knowledge are a great fit for such conferences. The [Journal of Online Trust & Safety](#) is the first journal of its kind to explicitly focus on T&S and plays a critical role in the T&S research ecosystem. It should continue to expand and, we hope, welcome other peer publications that collectively comprise richer academic literature for the burgeoning T&S field. T&S-related research should also continue to be published in journals focused on other academic disciplines that partially overlap with T&S.¹ This is especially true for specific harm areas. For example, the T&S and cybersecurity communities writ large still generally fail to reference terrorism studies literature, despite the fact that that field has been writing about risks online for more than twenty-five years.

The T&S vendor community would ideally continue to mature and find its voice and role in the larger ecosystem. Some existing vendors are already playing important roles in supporting [convening](#), [community building](#), and [education](#) programs—establishing an important precedent as early leaders in the space. The development of the cybersecurity vendor community pushed the industry toward greater investment, publications, benchmarking, and competitive progress, albeit sometimes at the expense of other dimensions (e.g., threat inflation, overcomplicating technological concepts, etc.). While RSA is a bit overwhelming and has a completely different zeitgeist and purpose than DEF CON (described below), certain cyber vendors do contribute to substantive security and policy activities, and push the field in a good direction. Other vendors are more extractive and prioritize their business imperatives over broader contributions to the field.

Finally, it would be remiss not to highlight the immense role that hackers have played in helping structure the cybersecurity field, driving innovation, transparency, and research forward. Who are or will be tomorrow's T&S hackers, and how can one ensure that the field will also benefit from outside (and, frankly, adversarial) perspectives? How can T&S integrate the depth of practical expertise in adjacent civil society, law enforcement, journalism, and research communities and channel the positive elements of hacker culture and community? Cultivating an unambiguous and grassroots T&S community culture—along with sophisticated vendors—will be key steps in the maturation of the T&S ecosystem.

NARRATIVE AND STORYTELLING

The cybersecurity community [has struggled to connect](#) with mainstream audiences and make its narratives accessible to nonexpert communities, deferring instead to storytelling—whether word or image-based—centered around threats and jargon that disempower users. [Studies](#) have shown that cybersecurity imagery focuses on locks, men in hoodies, and other visuals that do not communicate cybersecurity in any meaningful way or help users identify what they can do to stay safe. For that reason, the Hewlett Foundation funded a global contest to create new, more inclusive [cyber visuals](#) that are openly licensed for use and convey the complexity and reach of cybersecurity.

Another source of tension comes from the misallocation of the security burden. In the current ecosystem, the cybersecurity burden rests almost entirely on end users, who are often blamed for poor security outcomes.

¹ Journals serving various other, mature and nascent fields have published T&S-related research for years, including: Internet governance, cybersecurity, Internet Policy, Internet freedom, platform governance, HCI, online terrorism and violent extremism, disinformation studies, online forensics, STS, communications, political science, and security studies.

While some organizations are indeed negligent in their cybersecurity practices, most organizations lack the knowledge or capacity to improve their posture. Implementing effective cybersecurity requires considerable time and investment; moreover, since cybersecurity “standards of care” are not clearly defined for most industries, end users have difficulty knowing if they have invested enough in cybersecurity. Only in recent years has the “blame-the-user” narrative begun to shift to a secure-by-design approach that instead emphasizes the unique need for large platforms/providers to take responsibility for safeguarding users.

Learning from cybersecurity’s example, T&S will benefit from focusing much earlier on—identifying and clarifying for external audiences what T&S is, what success looks like, and why it matters, including through clear visuals and systems maps. Right now, users of platforms know what bullying or disinformation is, but lack an understanding of the role of the T&S field, how it does its job, etc. A mix of both written words, static images, and multimedia elements are necessary to redirect parts of the conversation around online harms to a conversation about the T&S ecosystem and how it can be leveraged for solutions. T&S must also narrate its positive benefit and opportunities for prosocial engagement, rather than solely focusing on harms, risks, and negative elements of the online experience. Focusing too much on harms and downside risk can feed into the perception of T&S as a lost cause or cost center (not deserving of additional investment).

Cybersecurity also has benefited from the [evolution](#) of an expert cadre of cybersecurity-focused [journalists](#). Numerous beat reporters have carved out a successful cybersecurity focus, and reporters covering national security, business, and other areas have also successfully reported on the role of cybersecurity within those fields. The best of these journalists have contributed to balancing out media coverage to make it more educational and not as fear-driven.

Within T&S, a growing number of journalists are helping build media expertise with the field, but they are heavily concentrated within the United States and focus almost exclusively on the major social media platforms. Reporters play a critical role in educating decision-makers in government (and elsewhere) about the nuances of T&S issues, explaining the importance of properly resourcing T&S work, and identifying where T&S needs have been dismissed or undermined. Building the field of reporters who can cover T&S, as well as the field of local reporters who can shed light on harms and risks for different communities (particularly marginalized communities or individuals in emerging markets), will be critical to moving broader T&S objectives forward and right-sizing the T&S community. It is critical for journalists to build relationships with T&S experts and civil society experts to inform their reporting; relying solely on industry voices risks imbalanced reporting and skewed narratives. Academic fellowships for T&S-focused or -interested journalists modeled on those at the [Alperovitch Institute](#), and focused events (such as [Verify](#)) could also support journalists’ knowledge development/education, just as they have within the cybersecurity field. Finally, industry will benefit from a more mature approach to interacting with reporters on T&S questions, engaging not only transactionally or defensively, but also with an eye toward building long-term, substantive relationships.

INFORMATION SHARING, IDENTIFYING THREATS, AND MEASURING HARMES

Information sharing has taken a long and winding path in cybersecurity. A mix of corporate opposition to sharing mandates, legal concerns about [antitrust liability](#), lack of [trust](#) in peer institutions and government partners, and other dynamics caused a series of legislative fits and starts before legislation was finally enacted in the United States in 2015.

As a leading [article](#) explained:

The theory behind . . . information sharing is clear and uncontroversial, even if the details of what to share, how best to do it and who to share with may sometimes result in debate and disagreement. The theory goes that organizations are better off sharing information and improving situational awareness than trying to recognize and face . . . threats and challeng-

es on their own. Some collective and coordinated efforts can help to identify, learn about and fend off threats and would-be attackers—as compared to acting individually with less information and situational awareness. That is also a reason why armies gather intelligence, where feasible, before going to battle.

Sharing information about . . . threats, incidents and vulnerabilities has some similarities to the concepts of a “neighborhood watch.” For both, the idea is to observe, gather and share information . . . to enable targets to recognize threats and defend better, reducing the likelihood that those attacks and attackers will succeed. In economic terms, we are seeking in part to raise the costs to attackers by using information sharing to shorten the time and narrow the instances in which their tools can be re-used profitably—as potential victims could develop defense tactics more quickly. To succeed as often, attackers would have to invest more in new or modified tools, or choose different targets—making it more expensive for them to generate each dollar in nefarious returns. We also seek to lower the cost of defense by helping defenders know what to look for and prioritize, and how to defend against those threats effectively.

Within cybersecurity, various [forms](#) of information sharing have evolved over time and can help provide inspiration and ideally faster piloting and iteration. These forms range from informal exchanges among practitioners to formal interorganizational mechanisms, such as [Information Sharing and Analysis Centers \(ISACs\)](#). Most industries have created an ISAC to collect, analyze, and disseminate actionable threat information to members and provide them with tools to mitigate risks. Certain more mature, well-resourced and high-risk industries, such as financial services, have taken this approach, creating, for example, the Analysis & Resilience Center (ARC) “to proactively identify, analyze, assess and coordinate activities to mitigate systemic risk to the US financial system from current and emerging cyber security threats through focused operations and enhanced collaboration.”

It is critical for T&S to learn from this experience for a few reasons. First, information sharing will likely be even more politically fraught within T&S than it is within cybersecurity. Information in T&S not only includes metadata and other (nonprivacy-invasive) adversary tactics, techniques, and procedures (TTPs), but also personally identifiable information such as account names, behavior, and content. Such content is much more closely regulated under the European Union’s General Data Protection Regulation and other privacy laws. A fair number of cybersecurity breaches deal with data only (ransomware attacks on hospitals or spyware surveillance of activists are troubling exceptions). But if a cyberattack disrupts the electric grid and someone dies because their oxygen machine stops working, that is a significant harm. That is arguably both a cybersecurity and a safety harm at the same time.

This leads to a second reason why information sharing improvements are critical: while cybersecurity failures primarily produce financial harm, T&S failures can result in acute physical [harm](#) or [death](#) on a [regular](#) basis. Given that, T&S should carefully and transparently address the challenges in linking trust (and all related information integrity and technology abuse issues) with safety (and all related mental and emotional abuse issues alongside material threats to physical safety). The specificity of harms across those categories, and their evidence on the face of limited information, differ. The relative practices and tradeoffs are most complex when both trust and safety are truly bridged by compound threats. Cybersecurity strove to resolve a similar challenge through the development of the [Common Vulnerability Scoring System \(CVSS\)](#). While the efficacy of the CVSS remains a contested issue within the cybersecurity field, having a framework that can help create a common definition of harms, their characteristics, and severity is a strength—one that would benefit T&S by providing clearer channels for information sharing across platforms and within the broader T&S community. The field also can learn from the financial services industry and how it has developed measurements of harm from malicious activity (including mapping monetary losses against the cost of cybersecurity investments).

Meaningful progress on information [sharing](#) within T&S will require meaningful investment and research, but there is a foundation to build upon given existing and nascent sharing with respect to [perceptual hashes](#) (i.e., unique digital representations for content) in child safety, violent extremism, and the sharing of nonconsensual intimate imagery, etc. Competition among and between certain companies (e.g., certain social media platforms) may undermine cooperation, however.

Within the T&S field, information sharing is nascent and most established in the [child safety](#), [violent extremism](#), and counterdisinformation spaces. The time has come for T&S to work through the thorny privacy and legal issues to develop a clear blueprint for one or more ISAC-like organizations. It is worth noting that cybersecurity has benefited from being a regulatory area that—at least in the United States—can support governmental alignment with industry, end users, and other stakeholders on cybersecurity adding value across the board. In addition, many governments have invested heavily in training people, developing policies, creating organizations, and passing legislation dedicated to cybersecurity. All of this activity smooths pathways to effective information sharing, aligns normative standards, and deepens a collective lexicon.

Notably, cybersecurity is also a field where state-led action and agreements have remained inaccessible and opaque to a broader community of stakeholders. National security and cybersecurity claims have frequently shielded contracts from scrutiny or oversight, and have also been used as a pretext to bar civil society, researchers, or journalists from accessing information regarding critical decisions or documentation of the activities being conducted in the name of cybersecurity. The T&S industry can learn from this example by building and protecting transparent (or at least not entirely opaque), multistakeholder processes from the outset as a de facto standard for the field.

Finally, the cybersecurity community has developed sophisticated methodologies for characterizing vulnerabilities and malicious activity. The CVSS provides a standard way to document the principal characteristics of a vulnerability and produce a numerical score reflecting its severity that can then be cataloged. The [Exploit Prediction Scoring System](#) (EPSS) provides an estimate of the likelihood malicious actors will exploit a given vulnerability in the next thirty days. These systems complement the [MITRE ATT&CK](#) framework, which is a globally accessible knowledge base that feeds into the development of specific threat models. In addition, cybersecurity has developed best practices around various methods of security [disclosures](#) and even [bug bounty programs](#), which “offer monetary rewards to ethical hackers for successfully discovering and reporting a vulnerability or bug to the application’s developer” as well as other nonremunerative disclosure mechanisms. The T&S field would benefit from [adapting](#) the concept of security disclosures, including bug bounties, to disclose both “vulnerabilities” in policies and enforcement. This would create an avenue for collaboration and discussions, as well as for companies to reward and incentivize good faith collaboration from academic researchers and individuals alike.

SCRUTINY, POLITICIZATION, AND RISK TOLERANCE

While the cybersecurity field has grown more capable over the past two decades, it has still failed in many respects to earn and maintain users’ trust. Lack of trust stems from several problems, including the fact that large-scale breaches remain commonplace, often due to companies failing to follow best practices. The recent scourge of [ransomware](#) is a case in point.

A cyber insurance industry has developed to help tackle some aspects of cybersecurity [risk](#), but to date cyber insurance has not driven companies to improve their cybersecurity as much as policymakers hoped. Insurers initially took many policy holders’ self-reported security and practices at face value, which often proved wrong or exaggerated. This approach is changing as ransom payouts have become unsustainable for insurers’ bottom lines. Insurers are now applying more rigorous criteria for issuing policies and making payments, as well as demanding more evidence regarding a company’s cybersecurity practices, which may drive companies to increase their cybersecurity investments. Despite this turmoil, companies often do not

suffer material, [long-term](#) financial consequences from subpar cybersecurity practices. This fact reduces companies' incentives to invest in cybersecurity.

T&S is in an even more difficult situation. This field is under much more scrutiny and is already becoming quite [politicized](#). In that respect, there are parallels with the threat intelligence research community, which has also experienced highly political adversarial attention (e.g., due to attributing cyberattacks to Russia, China, etc.). Within the field, politicization also includes bullying and [harassment](#) of T&S staff and academic researchers in an effort to influence their behavior and chill their speech. This troubling trend will impact the field for years to come, and it is critical to get ahead of this problem before it's too late through clearer approaches to establishing protection for those working in the T&S field and its broader ecosystem.

T&S workers will require additional training and resources to safeguard themselves from malicious actors who seek to harass, intimidate, and otherwise compromise them. These risks have long existed for cybersecurity experts—but in a less politicized environment. T&S, sadly, will need even more support to contend with the bad faith lawsuits and harassment campaigns that have already begun. Companies must dedicate additional resources to safeguard not only their T&S leadership, but all T&S staff who are at risk, and philanthropies will need to step up to fund protections for academic and civil society experts who face these same risks.

With regard to risk tolerance, media coverage of T&S issues has proven a valuable lever to generate attention for certain harms caused by misuse of platforms or products, but it also can be used to exaggerate edge cases (i.e., those occurring at the extremes of operating parameters) and make them a company's focus for attention and resources even when other harms are arguably more widespread and producing broader impact. These dynamics lead to public relations-driven investments in T&S (e.g., corporate responses to a damaging news cycle) as opposed to strategic investments in addressing the most acute risks/harms. Resources will be allocated to rare but public problems, rather than the most omnipresent problems because the common challenges have not made for a sensational news story.

This dynamic is exacerbated by the lack of any current shared or standard understanding of risk tolerance within T&S. It also negatively impacts the coherence and sustainability of in-house T&S efforts, and can undermine building solutions for pressing but less visible T&S challenges. Very few (if any) companies have defined what acceptable levels of T&S failure are and how to measure them. Companies are still struggling to measure risk across their systems, including dependencies between and among companies. In the absence of such basic rubrics, any public story about a T&S failure has the potential for major (negative) impact.

This is another reason why the T&S community must urgently craft a framework for defining harm, establishing acceptable levels of it, and defining how it is measured. How to define levels of acceptable failure will no doubt be challenging and require input from leading practitioners, academics, and policymakers (perhaps inspired by the Asilomar Conference for biotechnology in the 1970s). Such a framework would give companies a consistent method to allocate resources in response to news stories and/or activist complaints. Resisting the urge to treat every bad news story as a crisis will remain challenging, but a consistent, quasi-empirical basis for responses would improve broader efficacy with T&S. Best practices are urgently needed in T&S to assess vulnerabilities for public disclosure as well.

Advocacy and journalistic communities will also be crucial to building stronger technical understandings of how underlying services operate, and what the driving incentives of those services are. For example, many civil society organizations have found ways to build trust with companies, working collaboratively to fix a problem or address a harm prior to going public with their concerns. This is one of many reasons that it is critical for companies to build more effective pathways for engaging with and supporting external experts to help further T&S outcomes, drive broader progress, and shape narratives that allow for constructive engagement from a broad range of stakeholders. The lack of strong working relationships between companies and outside experts and activists risks unnecessary conflict, distraction, and further misallocation of resources.

DECISION-MAKING, LEADERSHIP, AND SUCCESS

Twenty years ago, cybersecurity roles were ill-defined, as was the path to becoming a chief information security officer (CISO). Refined concepts of cybersecurity governance (e.g., who is responsible for what with respect to cybersecurity) and working cross functionally have only recently taken hold (in some larger and more sophisticated organizations). So, too, has cybersecurity built out scalable team structures with clearer goals, targets, and objectives and key results (OKRs), setting teams up to more effectively drive business decisions and work cross functionally within companies. Cybersecurity also has made meaningful strides in how it should be incorporated into products and services. It is well understood that security can no longer be a post hoc, simple attachment, but should be a key attribute that needs to be designed into the base product (e.g., via security-by-design processes). This change in the product life cycle is a work in progress and faces opposition, especially where companies are motivated to be the first to market (e.g., GAI, etc.).

T&S by contrast still cannot define “what good looks like.” This lack of a basic understanding of good or successful T&S is where the cybersecurity field was two decades ago. Poor understanding of the trust and safety stack contributes to this lack of a North Star as does the lack of a clear governance framework. Likewise, the field needs to build on the work of the [Trust and Safety Professional Association](#) to not only map [potential](#) organizational structures for T&S teams, but also to identify which structures best fit differently-situated organizations.

Indeed, maturity models are lacking throughout the field. Whereas in cybersecurity “owner/operator” is a clear paradigm, T&S struggles to articulate an analogous governance model. Connecting T&S harms/successes to business impact—and measuring those—is another large gap. Finally, T&S must find a path to cross-functional influence, which is tricky given its different origin points (e.g., operations, compliance, customer service, etc.). Typically, these verticals can be less influential than the engineering origins of cybersecurity, which improved cybersecurity’s ability to establish cross-functional influence within many organizations. One focus in T&S should be on standardizing safety-by-design and graceful degradation as a norm across all companies. This is one of the most promising ways to ensure T&S equities are always considered and can be addressed in a timely fashion should risks be identified before a product is released to users.

COUNTERBALANCING GLOBAL NORTH DOMINANCE

The Global North has long dominated the cybersecurity field. While Global Majority representatives play an active role in certain high-profile [commissions](#) and at the [United Nations](#), they do not drive the allocation of resources globally. That’s because most of the major companies with top tier cybersecurity capabilities are based in the United States, Europe, and East Asia, and more security research and investment happens in those regions.

Those involved in T&S should work to avoid these dynamics and ensure the nascent field is more globally balanced and inclusive. This is particularly important given that the majority of most large platforms’ users reside in Global Majority regions even if the platforms themselves are based in the northern hemisphere. Moreover, the harms suffered due to T&S failures impact the global majority (as well as marginalized communities in the Global North) most acutely.

Finally, there is a growing effort for wealthier, northern countries to allocate resources to support cybersecurity training and capacity building in the Global Majority. A parallel effort in T&S is urgently needed, too.

LOOKING AHEAD TO GENERATIVE AI

GENERATIVE TECHNOLOGIES AND THE INDUSTRY OUTLOOK FOR TRUST AND SAFETY

Generative AI refers to powerful algorithms that can produce or generate text, images, music, speech, code, or video.² These algorithms rely on large language models (LLMs), consisting of vast artificial neural networks and are trained by consuming and processing large amounts of data. While not a new technology, the wildly popular release of [ChatGPT](#) and [DALL-E](#) at the end of 2022 catapulted GAI and LLMs into the public sphere. Leading technology companies ranging from Google to Microsoft to [newer entrants](#), such as OpenAI and Anthropic, are investing heavily in developing their own LLMs and associated products for public use. Governments, investors, and innovators alike have refocused their attention on these models and the products they power given GAI's potential to reshape society. Policy, practice, business models, and threat models for GAI have been changing by the day since ChatGPT was publicly released in November 2022. While it is not clear how this technology or its use will evolve, it is clear that its impact will be transformational, and it is possible to forecast some ways in which it could be leveraged—positively and negatively—within the T&S ecosystem and particularly with regard to content moderation.

GENERATIVE AI: FRIEND OR FOE TO CONTENT MODERATION?

Generative AI changes the nature of influence operations online and the moderation of illicit content by reducing the financial cost, time, and technical expertise required to produce mass amounts of hyperrealistic harmful content and potentially spread it at scale. Automating the production of [fraudulent content](#), misinformation, spam, influence operations, and other forms of illicit online behaviors through GAI results in content that is [more convincing](#) than previous forms of misinformation. While the content produced by GAI is not always perfect, it is more difficult for consumers to differentiate real from fake content when produced by GAI rather than less sophisticated methods. Moreover, this increased volume of deepfakes not only risks flooding trust and safety systems with exponentially greater quantities of content that will need to be monitored, but also injects greater quantities of hard(er)-to-detect forms of high quality (and potentially harmful) fake content into the system, too.

LLMs may also change the nature of influence campaigns. Previously, information operations focused on easier-to-generate artifacts—text and image—but we have yet to see what a targeted disinformation campaign might look like in the era of easily developed video and voice content. Put simply: people do not yet have the reflex for critical consumption of video and images as they have for online text-based content.

While GAI drastically changes the scale and speed at which malicious online behaviors occur, it might also serve as a tool for trust and safety professionals looking to mitigate these very same harms. **There are three areas in which these models can help identify and mitigate the harms they introduce:**

1 Data curation

These models can be used to identify harmful and falsified content and scale human review, as well as identify particularly egregious and harmful content, limiting harms to content moderators (who otherwise must review them).

² For a deeper analysis of this topic, see [Annex 1: Current State of Trust and Safety](#); [Annex 2: Building Open Trust and Safety Tools](#); and [Annex 4: Deconstructing The Gaming Ecosystem](#).

2 Model training

There are multiple emerging techniques (e.g., [reinforcement learning from human feedback \(RLHF\)](#) and [constitutional AI](#)) to improve the output of these models as well as identify places to create guardrails against inappropriate use.

3 Post-deployment

Evaluation of existing content is a clear use case.

For example, GAI could help scale the evaluation of questionable or inaccurate information. Developers can now produce tools that can combat automated influence operations, such as browser extensions and mobile applications that automatically [attach warning labels to potential generated content and fake accounts](#), or that selectively employ ad-blockers to demonetize them. As [suggested](#) by Georgetown University's Center for Security and Emerging Technology, OpenAI, and the Stanford Internet Observatory, websites and customizable notification systems could be built or improved with AI-augmented vetting, scoring, and ranking systems to organize, curate, and display user-relevant information while sifting for unverified or generated sources.

As content moderation is highly labor intensive and LLMs are [equipped to follow a set of instructions](#), trust and safety professionals may be able to benefit from the [application](#) of GAI in combating large-scale spam, fraud, and influence operations. AI-powered content moderation could also facilitate analyzing user interactions quickly to reduce the risks of hate speech, bullying, or cheating (in a game), and potentially doing so while limiting front line staff exposure to toxic material and minimizing privacy risks to the user.

Generative AI [may](#) also offer unique potential to improve the quality of classifiers, especially in minority languages. For example, it could be used to generate synthetic data in various languages, label that data, and/or train classifiers all in a matter of hours instead of weeks or months. Those classifiers could be regularly tuned and updated and widely shared, thereby providing a [powerful tool](#) to trust and safety teams. A lack of diversity in image datasets that train these models can be [mitigated by creating synthetic data](#).

However, GAI alone will not solve all product integrity issues. Toxicity and abuse online are not simply matters of content-based harms, but can also involve highly nuanced [actor- and behavior-based](#) challenges, which current LLMs may be less equipped to solve. Furthermore, [LLMs are sycophantic](#), and have no internal model for truthfulness of factuality, and systems deployed today also do not learn in real time: training on data is up to a cutoff point due to the time-consuming nature of training. As a result, GAI is well suited to automate or assist with more static tasks but will struggle to pick up on ever-changing social contexts.

Automating content moderation through large language or multimodal models will require robust human monitoring and auditing to ensure models do not possess unexpected bias. Models must be regularly trained and realigned as company content policies change. There are additional privacy risks in AI-powered harvesting of content, especially as companies collect and store more user data and expand red-teaming exercises to include an ever-widening array of individuals (thereby increasing the risk of leaks and abuse of data, and LLMs, etc.).

It is not yet clear how GAI may impact the efficacy of broader technological solutions across the T&S ecosystem. For example, even as multiple countries consider requiring companies to use age-verification technologies as a form of child protection, GAI experts warn that tools relying on audio or video to prove identity may be rendered obsolete. Tools to identify and watermark synthetically created (or altered) content are already in development and could play a powerful role in helping consumers and businesses demand specific standards and safety measures for the use of synthetic media.

CONCLUSION

As T&S develops into a field that can engage more intentionally and constructively not only with its own practitioner base, but also with a wider community of experts, it will be important to remain thoughtful, purposeful, and efficient whenever possible. Looking to other industries and their evolution can save years of trial and error, and focus collective efforts and investments on the moves most likely to have the greatest impact. Preparing the field to evolve in an expedited fashion will also be crucial for proactively taking on emerging technologies and identifying the risks and opportunities they pose to broader goals of safety, dignity, and trust across online spaces. Leveraging what cybersecurity has learned as it has evolved as a field—while balancing immediate challenges and opportunities from GAI—will no doubt stretch the nascent T&S community’s bandwidth, but holds promise, too.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This annex reflects contributions from the following members of the Task Force for a Trustworthy Future Web: Eli Sugarman, Schmidt Futures; Michael Daniel, Cyber Threat Alliance; Camille Francois, Niantic; Dr. Rumman Chowdhury, Berkman Klein Center; Dave Willner, OpenAI; and Yoel Roth, UC Berkeley. It also reflects contributions from Contributing Experts Trey Herr and Safa Shahwan Edwards, Atlantic Council; as well as Brian Fishman, Cinder.

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.

ATLANTIC COUNCIL BOARD OF DIRECTORS

CHAIRMAN

John F.W. Rogers*

EXECUTIVE

CHAIRMAN EMERITUS

James L. Jones*

PRESIDENT AND CEO

Frederick Kempe*

EXECUTIVE VICE CHAIRS

Adrienne Arsht*

Stephen J. Hadley*

VICE CHAIRS

Robert J. Abernethy*

C. Boyden Gray*

Alexander V. Mirtchev*

TREASURER

George Lund*

DIRECTORS

Todd Achilles

Timothy D. Adams

Michael Andersson*

David D. Aufhauser*

Barbara Barrett

Colleen Bell

Stephen Biegun

Linden P. Blue

Adam Boehler

John Bonsell

Philip M. Breedlove

Richard R. Burt

Teresa Carlson*

James E. Cartwright*

John E. Chapoton

Ahmed Charai

Melanie Chen

Michael Chertoff

George Chopivsky*

Wesley K. Clark

Helima Croft*

Ankit N. Desai*

Dario Deste

Lawrence Di Rita

Paula J. Dobriansky*

Joseph F. Dunford, Jr.

Richard Edelman

Thomas J. Egan, Jr.

Stuart E. Eizenstat

Mark T. Esper

Michael Fisch*

Alan H. Fleischmann

Jendayi E. Frazer

Meg Gentle

Thomas H. Glocer

John B. Goodman

Sherri W. Goodman*

Jarosław Grzesiak

Murathan Günal

Michael V. Hayden

Tim Holt

Karl V. Hopkins*

Kay Bailey Hutchison

Ian Ihnatowycz

Mark Isakowitz

Wolfgang F. Ischinger

Deborah Lee James

Joia M. Johnson*

Safi Kalo*

Andre Kelleners

Brian L. Kelly

Henry A. Kissinger

John E. Klein

C. Jeffrey Knittel*

Joseph Konzelmann

Franklin D. Kramer

Laura Lane

Almar Latour

Yann Le Pallec

Jan M. Lodal

Douglas Lute

Jane Holl Lute

William J. Lynn

Mark Machin

Marco Margheri

Michael Margolis

Chris Marlin

William Marron

Christian Marrone

Gerardo Mato

Erin McGrain

John M. McHugh

Judith A. Miller*

Dariusz Mioduski

Michael J. Morell

Richard Morningstar*

Georgette Mosbacher

Majida Mourad

Virginia A. Mulberger

Mary Claire Murphy

Edward J. Newberry

Franco Nuschese

Joseph S. Nye

Ahmet M. Ören

Sally A. Painter

Ana I. Palacio

Kostas Pantazopoulos*

Alan Pellegrini

David H. Petraeus

Lisa Pollina*

Daniel B. Poneman

Dina H. Powell McCormick*

Michael Punke

Ashraf Qazi

Thomas J. Ridge

Gary Rieschel

Michael J. Rogers

Charles O. Rossotti

Harry Sachinis

C. Michael Scaparrotti

Ivan A. Schlager

Rajiv Shah

Gregg Sherrill

Jeff Shockey

Ali Jehangir Siddiqui

Kris Singh

Walter Slocombe

Christopher Smith

Clifford M. Sobel

James G. Stavridis

Michael S. Steele

Richard J.A. Steele

Mary Streett

Gil Tenzer*

Frances F. Townsend*

Clyde C. Tuggle

Melanne Verveer

Charles F. Wald

Michael F. Walsh

Ronald Weiser

Al Williams*

Maciej Witucki

Neal S. Wolin

Jenny Wood*

Guang Yang

Mary C. Yates

Dov S. Zakheim

HONORARY DIRECTORS

James A. Baker, III

Robert M. Gates

James N. Mattis

Michael G. Mullen

Leon E. Panetta

William J. Perry

Condoleezza Rice

Horst Teltschik

William H. Webster

**Executive Committee Members*

The Atlantic Council is a nonpartisan organization that promotes constructive US leadership and engagement in international affairs based on the central role of the Atlantic community in meeting today's global challenges.

© **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
www.AtlanticCouncil.org