



SCALING

TRUST

ON

THE

WEB

EXECUTIVE 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.

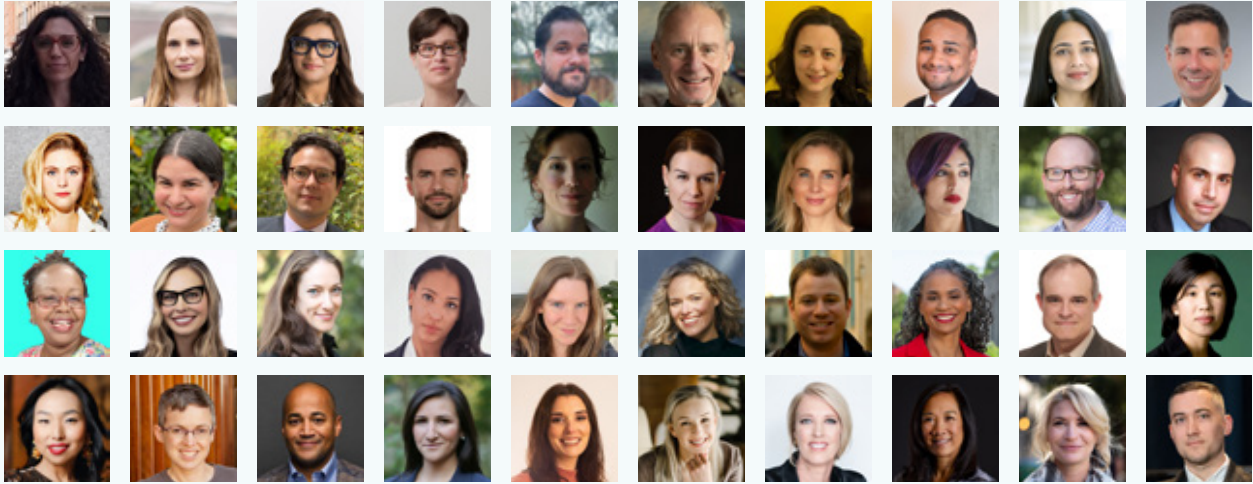
By looking beyond any particular challenge to the incentive structures defining—and constraining—the construction of our collective digital future, this task force has clarified where gaps in understanding or incentives must be addressed to further important change in building safer online spaces. Critically, 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

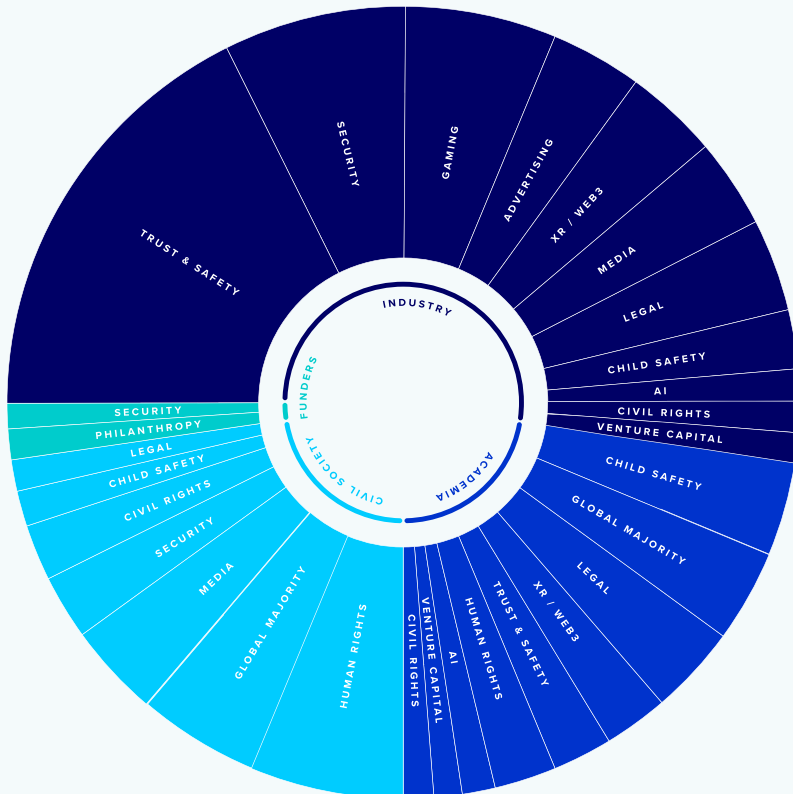
The task force comprises forty experts across industry, civil society, academia, and philanthropy. Every task force member brings deep expertise in at least two (and often three or four) of the following areas: policy development, AI, trust and safety, advertising, gaming, civil rights, human rights, 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.



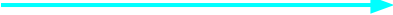
EXPERTISE OF THE TASK FORCE MEMBERS

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
Introduction	8
KEY FINDING 1 The Emergence of a Trust and Safety Field Creates Important Opportunity	12
New Initiatives Are Shifting T&S from a Community of Practice into a Field	12
Accelerating Knowledge Sharing About T&S Practices Is a Critical Need	13
Building Openly Available Tooling Is an Area of Opportunity	14
T&S Would Benefit from a Deeper and More Diverse Professional Pipeline	15
KEY FINDING 2 Academia, Media, and Civil Society Bring Crucial Expertise to Building Better Online Spaces	16
Academic Researchers Need Improved Access to Support T&S	16
Civil Society Expertise Is Crucial and Under Threat	17
The Media Is Fundamental to Improving Understanding and Accountability, and Also Under Threat	18
KEY FINDING 3 Protecting Healthy Online Spaces Requires Protecting the Individuals who Defend Them	20
KEY FINDING 4 Learning from Mature, Adjacent Fields Will Accelerate Progress	22
Cybersecurity	22
Human Rights	23
Additional Key Fields	24
KEY FINDING 5 The Gaming Industry Offers Unique Potential for Insights and Innovation	26
KEY FINDING 6 Existing Harms Will Evolve and New Harms Will Arise as Technologies Advance	28
Federated Spaces	28
Immersive Spaces	30
Generative AI	31
KEY FINDING 7 Systemic Harm Is Driven by Market Failures That Must Be Addressed	33
Measuring T&S Is a Meaningful Challenge	33
Emerging Regulation Is Already a Market Driver for T&S	34
The Role of Venture Capital Has Been Underexamined	36
KEY FINDING 8 Philanthropies and Governments Can Shape Incentives and Fill Gaps	37
Key Recommendations	38
Conclusion	45
Acknowledgments	46
Task Force Members	47
Expert Contributing Organizations	48
Contributing Experts	48

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.

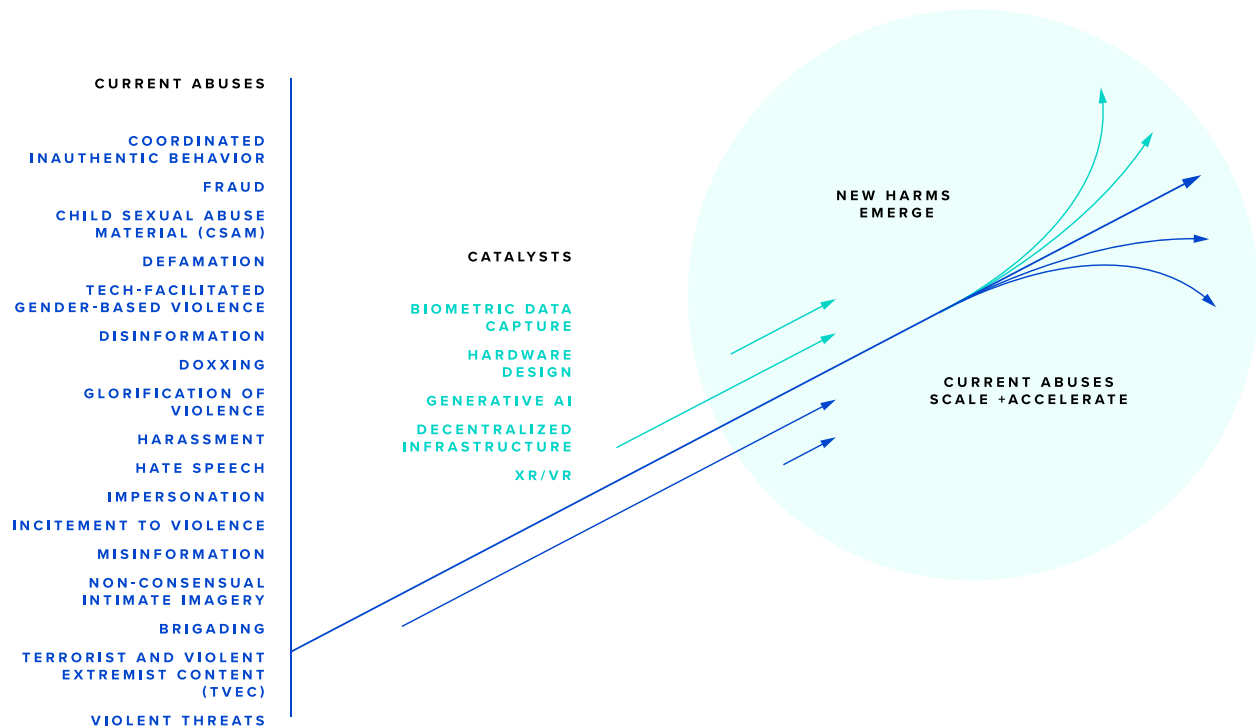
¹⁴ For a deeper dive into this topic, see *Annex 5: Collective Security in a Federated World*. Broadly speaking, the “fediverse” is a catch-all term for a wide array of distinct products, services, and platforms that interconnect using a set of shared communication protocols such as the W3C standard [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 on-line 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 Ogunnaike
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

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. Lodai

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