

ANNEX 6

SCALING TRUST ON THE WEB

LEARNING FROM CYBERSECURITY, PREPARING FOR GENERATIVE AI

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

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

ISBN: 978-1-61977-279-3

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

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

Please direct inquiries to:

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

For more information, please visit www.AtlanticCouncil.org

June 2023




ANNEX 6

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

TABLE OF CONTENTS

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

INTRODUCTION



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

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

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

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

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

LESSONS LEARNED FROM CYBERSECURITY'S EVOLUTION

EDUCATION, PROFESSIONAL TRAINING, AND RESEARCH

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

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

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

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

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

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

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

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

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

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

NARRATIVE AND STORYTELLING

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

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

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

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

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

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

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

INFORMATION SHARING, IDENTIFYING THREATS, AND MEASURING HARMES

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

As a leading [article](#) explained:

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

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

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

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

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

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

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

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

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

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

SCRUTINY, POLITICIZATION, AND RISK TOLERANCE

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

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

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

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

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

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

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

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

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

DECISION-MAKING, LEADERSHIP, AND SUCCESS

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

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

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

COUNTERBALANCING GLOBAL NORTH DOMINANCE

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

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

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

LOOKING AHEAD TO GENERATIVE AI

GENERATIVE TECHNOLOGIES AND THE INDUSTRY OUTLOOK FOR TRUST AND SAFETY

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

GENERATIVE AI: FRIEND OR FOE TO CONTENT MODERATION?

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

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

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

1 Data curation

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

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

2 Model training

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

3 Post-deployment

Evaluation of existing content is a clear use case.

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

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

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

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

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

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

CONCLUSION

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

AUTHORSHIP AND ACKNOWLEDGEMENTS

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

This report does not represent the individual opinion of any contributor, member of the Task Force, or contributing organization to the Task Force. Rather, it serves to consolidate collective research, feedback, and contributions gathered over a five-month period.