TRUST & SAFETY MARKET RESEARCH REPORT

A comprehensive business case for investment in the trust & safety industry
Bolstering the trust and safety (T&S) field has never been more important. T&S practitioners, teams, and vendors play a critical role in protecting users of digital technologies worldwide from the myriad harms they would otherwise encounter. They are society’s frontline digital defenders who must grapple both with the promise and peril of rapidly evolving technologies, especially generative AI.

Yet, not nearly enough is known about the incentives (or lack thereof) shaping the T&S market. What is clear, however, is that we are witnessing a market failure in which inadequate resources are being invested in the T&S community. This is not only bad for society – because it undermines safe online environments – but also bad for business – because consumers don’t want to be exposed to toxic content and behavior. I’m excited for this Trust and Safety Market Research Report because it grounds the conversation about T&S in business metrics. The report makes a compelling case for T&S by illustrating how it is a fast-growing discipline that contributes to business growth and innovation while simultaneously contributing to a more positive, prosocial internet.

Much like with cybersecurity ten or twenty years ago, now is the moment to focus on building and investing in the T&S field. Doing so is not only a charitable imperative, it’s also smart business.

Eli Sugarman
Former Cyber Initiative Director, William and Flora Hewlett Foundation
I founded Duco 7 years ago to empower leading companies to operate safely, securely, and responsibly by mobilizing the world’s leading experts to help solve complex challenges. Since we began, we have stayed true to that mission and have become a trusted resource for leaders from leading internet companies to solve some of their most pressing issues, including those related to trust and safety.

Throughout the past few years, we’ve heard from T&S leaders across the industry that while their function is critical to the success of their companies, it is often misunderstood by corporate executives, which prevents them from getting the investment needed to keep their respective platforms safe and healthy. At the same time, these problems are getting more, not less, challenging – especially in the era of artificial intelligence.

Moreover, while there has been growing investment into third-party T&S solutions, limited resources about this market potential lead to less informed and organized investment in the industry.

What follows is evidence that T&S fuels organic user growth, advertising revenue, and the overall bottom line. Now is the time to invest in T&S to ensure long-term success and solid returns on investment.

We’d like to thank the many T&S practitioners and vendors who inspired and informed the development of this report and are on the frontlines of this evolving and maturing industry. With this report’s publication, we hope that the T&S industry will see even more investment and recognition for the experts and solutions that maintain our digital environment’s safety.

Sidney Olinyk
Founder & CEO, Duco
# CONTENTS

**Executive Summary** ................................................................................................................. 4

**Market Overview** ....................................................................................................................... 7
  A. T&S Industry Overview ........................................................................................................... 8
  B. Market Segmentation ............................................................................................................. 10
  C. Market Size, Growth Trends, Projections, and Considerations ........................................... 14

**Why Invest in T&S?** .................................................................................................................. 20
  A. Capital Investments and Expertise .......................................................................................... 21
  B. Competitive Landscape and Market Players .......................................................................... 23
  C. Role of Emerging Technology and Opportunities in T&S ..................................................... 24
  D. Corporate Benefits of T&S .................................................................................................... 26

**Conclusion** ................................................................................................................................. 30
  A. The Time for Investment ........................................................................................................ 31
  B. Opportunities for Additional T&S Engagement ..................................................................... 32

**Appendices** ................................................................................................................................ 34
  A. Glossary of Key Terms .......................................................................................................... 35
  B. Methodology ......................................................................................................................... 37
  C. T&S Vendor Overview .......................................................................................................... 39
  D. Drivers of T&S Market Growth ............................................................................................. 54
  E. T&S Team Overview .............................................................................................................. 63
  F. Corporate Benefits of T&S Vendor Partnerships ................................................................ 65
  G. Case Studies .......................................................................................................................... 71
  H. Expert Investors .................................................................................................................... 77
  I. Data Sources and References ............................................................................................... 78
  J. Acknowledgements / Contributors ....................................................................................... 79
EXECUTIVE SUMMARY

More than 5.3 billion people are online – and those that aren’t are likely to be in the near future. Navigating online spaces is increasingly complicated and important, both for society and for the bottom lines of the companies leading the way in the online world. But the trust & safety (T&S) industry, while robust, still lacks the needed investment, standards, and recognition to properly address current and future issues. This presents an opportunity for investors and executives to lead the way and see significant returns by investing in a proven revenue driver – the T&S solutions market, which currently encompasses a total addressable market of $24.8B and a serviceable addressable market of $14.7B (see Appendix B - Methodology).

In a world where billions of people consume and create digital content and services every day, T&S solutions keep users happy and safe – which makes them stay on the platforms – while also protecting the companies. It’s a big task and one that has increasingly been under the microscope. T&S solutions broadly refers to the policies, processes, and digital tools that platforms employ to manage content and conduct related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety (DTSP) – efforts which foster safer platform environments by design.

Though T&S may currently be considered part of another department (such as legal, cybersecurity or customer experience) it must be recognized as its own unique function that is needed across all departments and funded as such. In recent years, there has been a surge in interest in the field following high profile news events, pressure from users and advertisers, and new regulations around the globe. That means digital platforms cannot ignore the critical role of T&S in the success of their enterprise. And T&S teams are on the frontlines of new demands for accountability.

T&S fuels organic user growth, advertising revenue, and the overall success of a company – and now is the time to invest in T&S to ensure long term returns on investment. Several emerging factors make investment in T&S solutions critical for both investors and company leaders:

1. **The rise of artificial intelligence** presents both a challenge and an opportunity for all companies with an online presence. Managing the risks of increased AI-generated content will be critical to success in this new digital frontier. But artificial intelligence technologies can also assist with these risks, such as using large language models to classify existing and new content trends.

2. **New laws and regulations** such as the European Union’s Digital Services Act which lay out rules for marketplaces, social networks, content sharing platforms and more – with hefty fines for breaking the rules.

3. **Macroeconomic concerns** have led to mass layoffs across digital platforms, but this provides an opportunity for emerging AI-enabled software solutions led by experienced founders to fill in the gaps in a $14.7B serviceable addressable market, with an average annual growth rate of 36.6%–57.2% over the next 5 years.
4. **News events** such as elections and conflicts around the globe will have all eyes on platforms for the foreseeable future.

Though there is room for significant growth in the industry, investment in T&S solutions is already a proven revenue driver. T&S investment within successful platforms, such as Airbnb developing a solo traveler safety feature and Minecraft deploying in-game player interaction controls, has demonstrated that this investment is a prudent business decision as it maintains healthy platforms and safeguards the brands. This is especially important in an era where 60% of U.S. adults now consider a brand’s values as a top factor when making a purchase. On the flip side, the divestment in T&S solutions and subsequent decline of daily users at X (formerly Twitter) is a cautionary tale proving the importance of the field (see Appendix G – Case Studies).

Platform executives and investors alike have a ripe opportunity to invest in this quickly maturing T&S industry, which will ultimately prevent, detect, and manage consequential digital conduct such as disinformation, fraud, harassment, hate speech, and violent extremism at a critical time in history.

Emerging factors in the market only make the opportunity clearer. The advent of generative AI (genAI) marks a transformative era that reshapes the T&S landscape. While the list of industries that employ T&S solutions has historically included consumer-facing platforms such as social media platforms, marketplaces, dating services, gaming companies, internet service providers, search engines, and video streaming services, the widespread adoption of genAI means companies in industries such as banking, defense, healthcare, retail, and life sciences will also need to deploy T&S solutions to manage the affiliated risks of AI-generated outputs. For instance, as healthcare clinicians leverage genAI for documentation, T&S solutions will need to be deployed to ensure the reliability of information being used. And for retailers using AI chatbots for customized shopping experiences, T&S solutions will need to be deployed to ensure chatbots are not providing false or misaligned information.

Currently, the outsourced T&S services category represents the largest portion of the $24.8B TAM at 56.8%, followed by in-house T&S software at 27%. This dynamic reflects that the biggest tech companies prefer to outsource services like human content moderation, but leverage their internal engineering resources for software development. However, T&S software solutions, facilitated by evolving AI capabilities, are poised to rapidly capture human-powered outsourced services market share at a yearly rate of 4% for the Tech Giants and 10% for the Enterprise and Mid-Market segments – growing the T&S software TAM from $7.4B to $15.4B, while the T&S services TAM only grows from $17.3B to $17.5B by 2028.
And as the market grows, so too will the availability of new tools to safeguard platforms and users. While the availability of generative AI tools is contributing to increased T&S issues, it is also revolutionizing software capabilities for T&S teams. The solution set has evolved over the past 5 years from focusing on manual processes (i.e. human content moderators), to turnkey products developed by T&S experts that emphasize automation and streamlined T&S workflows. Technological advancements, combined with an outflux of talent from large platforms like Meta, have led to the creation of highly effective solutions that can be deployed at scale and are ripe for investment. And experts predict that growth will continue alongside expansion of the field, with opportunities to deploy AI-enabled T&S solutions in more non-English languages and in all industries experimenting with AI-generated outputs.

While venture investors have been funding T&S solutions for over a decade, the past few years represent an inflection point. According to Pitchbook, from 2019 to 2023, the total amount of capital invested into the T&S software market alone, at $7.78B, was 4.7x the amount of capital invested from 2014–2018, and 40x the amount of capital invested from 2009–2013. Top firms have taken notice and are actively participating in this ecosystem. Funds like 8VC, Accel, a16z, Amplify Partners, Benchmark, Bloomberg Beta, Union Square Ventures, BoxGroup, Norwest, General Catalyst, and Thrive Capital have invested in the category.

In addition to an opportunity for investors, T&S solutions are also a proven revenue driver for companies. Adopting T&S practices directly impacts user churn, advertiser revenue, company reputation, and compliance management, and the growing T&S software industry is an ideal solution for key stakeholders looking to safeguard their brand. (see Appendix F – Corporate Benefits of T&S Vendor Partnerships).

This Report brings coherence to T&S as a distinct industry, which will facilitate the creation and adoption of new companies, jobs, policies, and multisectoral collaboration that can manage the sociotechnical risks that are affiliated with user- and AI-generated content and services. It also provides key decision makers with evidence as to why investment in T&S is synonymous with long-term value generation.
MARKET OVERVIEW

A. T&S INDUSTRY OVERVIEW
B. MARKET SEGMENTATION
C. MARKET SIZE, GROWTH TRENDS, CONSIDERATIONS
A. T&S Industry Overview

The early days of the internet were a time of hope and ambition for greater connection between individuals and communities across the globe, and lucrative businesses were developed during this “dot-com boom” with visions of grandeur and prosperity. Yet, many of these early builders did not create these businesses with the foresight of all the harms that may manifest from them. Instead they were developed with the promise of the possibilities that new digital spaces could generate – and the existing regulation frameworks encouraged this approach. But as individual blogs and websites scaled towards global, multibillion-dollar platforms and services, the affiliated risks concurrently grew. This gave rise to the cybersecurity industry in the 2000s (see Appendix G – Case Studies), when dedicated service providers were tasked with protecting the underlying networks, infrastructure, and data from evolving malicious attacks. These efforts have now grown into a lucrative and bustling cybersecurity industry, growing by $83B in annual revenue from 2016 to 2023, and resulting in many chief security officers and venture funds to manage and invest in these operations.

However, focusing on network security alone is not enough for these platforms – many of them also host content and communities that require additional oversight to ensure that overall platform health is aligned with their brand and standards. This is how T&S was developed. T&S manages content and conduct-related risks on digital platforms and services. While this field has recently seen a rise in professionalization (TSPA, DTSP, Integrity Institute), dedicated academia, and public recognition, tech companies have long relied on its efforts. Amid the rise of online commerce and communication in 1990s, the rise of the social web in the 2000s, major electoral events in the 2010s, and emerging technology formats in the present – where offline harms such as fraud, harassment, and child exploitation have new formats to proliferate within – T&S has grown into its own field. While the tools and teams deployed to manage cybersecurity and T&S risks may overlap, T&S is its own industry with a total addressable market of $24.8B – and requires dedicated analysis and investment to manage the evolving harms within its domain. To accurately measure the size of the industry and the opportunity, researchers for this report used Harmonic, a startup database that indexes 20M+ companies based on website information and filings, and identified 222 private sector companies that fall under the Report’s T&S solutions scope. In addition, researchers identified 261 platforms that could consume T&S solutions based on their core offerings and the presence of T&S employees on staff (see Appendix B – Methodology).

Before the professionalization of the T&S field, many platforms handled T&S issues in a bespoke manner, mostly due to legal, media, customer, and/or business pressures. For instance, earlier T&S issues managed on YouTube dealt with violent or extremist content, due to pressure from governments and advertisers. These issues were often managed under customer experience teams or as part of risk management, compliance, or site integrity. But as these conduct and content-related concerns grew, T&S developed as its own separate function. T&S as a field has evolved from creating community guideline policy and content moderation (rule setting and
enforcement) to a wide variety of functions involving research, design, operations, product, engineering, and data science to actively manage and mitigate content and conduct concerns within a platform’s processes and products (see Appendix E – T&S Team Overview). And in the AI era, those with T&S experience are being asked to design and lead safety efforts within AI companies themselves.

Yet, even as T&S teams grow in capacity beyond defining and moderating content, challenges regarding which unwanted content and conduct are allowed persist. And media pressure and increasing fragmentation of internet regulation continue. For instance, the announcement and release of the Digital Services Act required certain platforms to dedicate resources towards ensuring that they are complying with the expectations outlined for them by the European Union. In addition, the latest Freedom on the Net report highlights that at least 48 countries pursued new rules for technology companies in 2022, half of which relate to how platforms treat content.

The T&S industry is at a critical inflection point that is ripe for analysis and investment. The professionalization, organization, and definition of this field, requires investors and business leaders to understand why they need to set aside resources and investments to foster its growth. As a follow-on to the Scaling Trust on the Web report released in June 2023, this Trust & Safety Market Research Report provides insights and analysis to illustrate why there is not only a security and human rights case for T&S, but also a business case for it.

For the purposes of this Report, *trust & safety is defined as the policies, processes, and tools digital platforms employ to manage content- and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety* (DTSP), efforts fostering safer platforms by design. Though there is no industry standard definition of T&S, teams are aligned on their belief that their work is both mission and business critical for a platform to not only survive but thrive in an unpredictable world, as their efforts directly impact user churn, advertiser revenue, company reputation, and compliance management. But, just as there is a lack of a standardized definition of T&S, there is also a lack of standardized metrics used to communicate its impact within and outside of these platforms. Associations such as the DTSP are focused on developing standard terminology and metrics to align the field, so that T&S practitioners can better communicate why their work should be prioritized and resourced.

While standardizing industry definitions and metrics are certainly steps in the right direction, the T&S industry needs a comprehensive overview to demonstrate its role as a growth strategy, rather than a cost center. Because these teams are not viewed as revenue drivers, they are instead de-resourced and deprioritized by business executives, and by extension, investors who deploy capital for tools within the industry. T&S teams are some of the *first to be laid off*, even though they maintain platform health, boosting user satisfaction and retention. This Report addresses this fallacy and provides data-driven evidence as to why T&S investment creates long-term value generation for platform executives and potential investors. While there has already been an influx of venture capital investment into T&S software startups, a comprehensive resource is needed to support
both investors evaluating potential investments in this domain and executives who are considering deploying capital towards managing T&S efforts.

**B. Market Segmentation**

As the digital landscape evolves, T&S has been at the forefront of addressing challenges posed by managing content and behavior within emerging formats, including user or AI-generated text, video, audio, and XR/Metaverse. While industries that have historically used T&S include consumer-facing platforms such as social media, search engines, messaging applications, streaming services, gaming, dating apps, the sharing economy, web services, AI content generators, and app stores, the widespread adoption of genAI means all sectors, such as healthcare, financial services, education, news, hospitality, and retail, will also need to deploy T&S solutions to manage the affiliated risks of AI-generated outputs. Companies using generative AI, especially in regulated sectors, must rigorously manage foundation model risks to ensure reliable outputs and safe experiences for their customers.

Example platforms currently using T&S Solutions

While platform approaches to T&S vary, at a high level, activities that fall under the umbrella of T&S generally support one or more of the following workflows:

1. **Creating and maintaining policies**: Developing policies based on company and/or legal requirements for what content and/or conduct is and is not allowed on a platform.
2. **Detecting content and/or conduct** on the platform.
3. **Taking action**: Acting on problematic content and/or conduct.
4. **Measuring progress:** Measuring and tracking actions taken on content and/or conduct and the impact on key metrics.

5. **Sharing results:** Publishing transparency reports outlining actions taken and their outcomes, which includes reporting for regulators and law enforcement.

The solutions that companies adopt to support these activities, known as the T&S solutions stack, take on varying levels of complexity depending on the types of products that platforms support. (see Appendix C ~ T&S Vendor Overview.) Note that while many T&S teams also have personnel dedicated to consulting on how platforms should proactively build safe products, this Report is focused on the marketplace of vendors who provide support for the activities mentioned above.

**T&S Solutions Market Overview**

T&S solutions are designed to assist T&S teams that address a variety of risks, including but not limited to: copyright infringement, child sexual abuse material (CSAM), election interference, misinformation, fraud, harassment, nonconsensual intimate imagery, spam, synthetic media, terrorist and violent extremist content, and more. And with the surge of generative AI-produced content containing harms such as misinformation and illegal CSAM, this threat landscape is exponentially increasing.

A bottom-up analysis of 222 private sector organizations yielded 2 types of outsourced T&S solutions: **T&S BPO services** and **T&S software**. Within T&S software, the analysis yielded 2 market segments that are increasingly leveraging AI technologies to manage T&S risks: **SaaS moderation platforms** and **specialized tools**, which includes risk and threat intelligence, detection models and APIs, age and ID verification, and watermarking and information provenance.

**T&S Business Process Outsourcing Services (BPOs)**

Business processing outsourcing services (BPOs) are firms that platforms use to help with T&S policy enforcement/operational needs. While BPOs provide a spectrum of services, such as customer service management and enterprise modernization, content moderation or “content trust and safety” comprises a sizable part of the T&S BPO market.

BPO employees who are subcontracted for content moderation services fulfill a variety of functions, including but not limited to:

- **A.** Assess potentially harmful content for adherence to company guidelines and legal/reporting requirements
- **B.** Tag content with additional metadata to help train AI systems to detect content automatically
- **C.** Flag problematic content for further review
- **D.** Triage content to other teams
- **E.** Take action on violating content
While there are other types of outsourced services that T&S teams deploy (e.g. expert advisory services), such companies were not included in the analysis since they have smaller relative contracts to BPOs.

T&S Software

1. SaaS moderation platforms
T&S SaaS moderation platforms are purpose-built to specifically manage the complex operational issue of user-generated content (UGC) for T&S teams. Technological advancements have enabled these vendors to incorporate best in class third party AI detection models, APIs, risk intelligence features and other customized solutions to automate and streamline their customer platforms’ specific needs.

2. Specialized tools
T&S specialized tools are designed to provide insights that aid T&S teams in their efforts in one of the following domains:
   A. OSINT/risk and threat intelligence: Open source intelligence, risk intelligence, and threat intelligence tools collect, process, and analyze data to understand emerging hazards or trends.
   B. Detection models/APIs: Companies provide AI models and APIs to assist platforms in detecting specific types of content, ranging from audio, to video, and various types of written content.
   C. Account/ID verification: Products that verify user accounts or identities, including age verification.
   D. Watermarking/information provenance: Products that identify the origins of media online, used to help users distinguish between human, AI-generated, and/or edited content across audio, image, text, or video.

While these technologies may also be used by a variety of functions outside the realm of T&S responsibilities (e.g. cybersecurity, fraud management), they are included in the analysis since these capabilities serve a critical role providing expertise among various aspects of the T&S workflow.

Summary Statistics: T&S Solutions Vendors
Our sample was split between ~20% T&S services\(^1\) and ~80% T&S software. Although outsourced services providers make up a larger portion of the market and employ 96% of T&S solutions employees, the higher ratio of T&S software solutions reflects the growing early-stage investment into T&S software startups.

Additionally, approximately 35% of T&S solutions were identified as pure-play (T&S solutions is their core offering) and 65% include diversified providers that offer other products or services in addition

---

\(^1\) BPOs are not necessarily considered startups and so may be undercounted in the report, although ~20% of the sample does fall into this category.
to those in T&S\textsuperscript{2}. This can be attributed to the versatility of specialized tools, which made up 71% of the T&S software sample, as they may also be used by functions adjacent to T&S, such as fraud management and cybersecurity.

Vendors operate across the world, with U.S., E.U., and U.K.-headquartered providers dominating 52%, 25%, and 13% of the market respectively. The U.S.’ dominance in the T&S solutions market can be attributed to both the breadth of U.S.-based T&S talent and the availability of investment dollars relative to other countries. The E.U. and U.K.’s significant presence can be attributed to its leadership in regulation (i.e. Digital Services Act, Online Safety Bill) which has driven their government’s and investors’ commitment to support Safety Tech\textsuperscript{3} innovation.

However, there are few T&S solutions based in Asia, despite the vibrant startup ecosystems in the region. This may change as laws passed in 2023, such as the Digital India Act and Singapore Online Safety Act, may spur more innovation for T&S solutions.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{tssolutions_by_hq_location.png}
\caption{T&S Solutions by HQ Location}
\end{figure}

\textsuperscript{2} Pure-play companies were identified as those that have “content moderation”, “trust & safety”, “trust and safety”, “misinformation”, “disinformation”, and/or “fact check” in their company descriptions.

\textsuperscript{3} For more information on the U.K.’s approach, please see Paladin’s U.K. Safety Tech Report.
C. Market Size, Growth Trends, Projections, and Considerations

Of the T&S buyers analyzed in this report, the 15 largest platforms, dominated by Google and Meta, accounted for approximately $16.5B in spending annually, or roughly 66% of the $24.8B total addressable market, characterized by heavy reliance on BPOs. Some Tech Giants have publicly shared their T&S budgets in this era of heightened platform accountability. Ahead of the 2024 elections, Meta has spent more than $20B since 2016, and TikTok pledged to spend $2B in 2024 alone to protect users.

Generally, these companies generated revenue from 4-8 products featuring UGC or UGS (user-generated services) with user bases ranging from hundreds of millions to billions of accounts. The societal ubiquity of these platforms enables the large investments in T&S solutions that are necessary to maintain user experiences at scale and to prevent real-world harm that originates online. While smaller Enterprise and Mid-Market companies lack the scaled user base necessary to support such large T&S budgets, platforms with large young user bases, such as Discord, where more than 15% of employees are dedicated to user safety, have increased T&S investment in a new era of accountability.

T&S services account for $17.3B, or 70% of the total addressable market, while T&S software accounted for the other 30% of the TAM.
Thus, the outsourced services category represents the largest portion of the TAM at 56.8% ($14.1B), followed by in-house software at 27% ($6.7B). This dynamic reflects that bigger tech companies prefer to outsource services like human content moderation but leverage their internal engineering resources for software development. It also suggests that while T&S software machine learning models are improving, they have yet to outperform what human moderators can provide because of their cost and high implementation effort.

Tech Giants do the most outsourcing (primarily in the services category), while Enterprise and Mid-Market teams choose to keep a higher percentage of staff in-house. It appears that the largest companies benefit heavily from outsourcing economies of scale and take advantage of lower unit costs of larger, longer-term contracts with BPO services. Tech Giants tend to find little value in outsourced software since their software development and proprietary datasets are more developed and customized for their specific needs. Exceptions exist in cases where a vendor offers niche datasets or subject matter expertise in implementation and support.

However, Enterprise and Mid-Market firms are more likely to use software vendors because they lack the internal resources and data teams present at the Tech Giants, presenting an opportunity for investors.

**Drivers of Market Growth**
As platforms become more integrated into daily life, the marketplace of solutions that address affiliated T&S issues will continue to evolve. Since failing to address T&S can be catastrophic for
Platforms operating in complex digital environments\(^4\), experts anticipate that market growth will continue to accelerate, influenced by both business pressures and macroeconomic factors.

Platforms undertake activities related to T&S\(^5\) for the following business and social responsibility reasons:

1. **Maintaining brand safety**
   Platforms must be able to ensure that available advertising inventory is not associated with inappropriate content that could damage a brand’s reputation and risk the loss of advertising revenue. For example, X (formerly Twitter) lost approximately \$75M in advertising revenue after a 919%+ increase in antisemitic content (see Appendix G – Case Studies). Additionally, platforms must maintain their own brand integrity by upholding their community standards of acceptable content/conduct.

2. **Media scrutiny and increasing consumer expectations**
   Effectively managing content and conduct-related risks improves public perception and mitigates negative press coverage. Media scrutiny has also led to increased consumer expectations for an improved user experience with more robust T&S safeguards and quality content – if safety measures do not meet consumer expectations\(^6\), consumers have a rising number of platform alternatives to take their business and attention to.

3. **Improved industry standards**
   There is also more professionalism among T&S employees at platforms, who develop best practices and shared industry resources through organizations such as the Trust and Safety Professional Association, the Integrity Institute, the Fair Play Alliance (gaming), and Coalition for Trusted Reviews (e-commerce). This has prompted companies to pay attention to the investments the organizations are making, raising the bar for new entrants.

4. **Emerging regulation & compliance**
   Over the past decade, governments worldwide have been increasingly introducing legislation on T&S issues. And companies are required to comply with requests from law enforcement and regulations across the globe such as the EU AI Act, Digital Services Act (EU), Online Safety Act (U.K.), Online Safety Act (Singapore), and Australia’s Online Safety Act (see Appendix D – Drivers of Market Growth). As platforms increase their global scope, companies will need to equip their T&S workflows with the appropriate localization to prevent potential fines.

---

\(^4\) Formspring, an early success of the anonymous online Q&A genre with 1M users within 45 days and $14M raised, hosted hateful content that led to a series of teen suicides and subsequent boycotts. While the site eventually added features designed to discourage bullying, the site shut down within 4 years.

\(^5\) A workshop with T&S professionals and academics found that most platforms heavily weight content quality metrics in their algorithms to increase long-term user retention.

\(^6\) As a result of prioritizing posts containing “knowledge and advice”, LinkedIn has seen a 80% reduction of user complaints about irrelevant content on their feeds.
Macroeconomic factors will also influence growth:

1. **Geographic expansion**
   Revenues from digital advertising (social media and other channels), e-commerce platforms (ride-hailing, delivery), and other mobile applications (in-app purchases and paid apps) underpin T&S budgets. Therefore, growth in the population of active internet users and their spending power will drive growth, especially as 59.6% of the global population actively uses the internet at present, with most non-internet adopters living in growth markets.

2. **Digitized society**
   Market penetration statistics indicate that the diversification of online platform applications and their increased importance in everyday activities will fuel the need for T&S solutions to manage these new mediums.

3. **Outflux of T&S talent**
   Founders are taking the nuanced expertise they honed addressing T&S issues at the major platforms and building new T&S software startups. A combination of T&S professionals noticing gaps in the T&S market coupled with technology industry layoffs have led to a major outflux of talent from companies with acute T&S issues and sizable budgets to address them at scale. Teams’ deep and practical understanding of the problem space should help them define the category and attain product-market fit.

However, there are some key considerations for the T&S market:

1. **Market concentration**
   While the SAM analysis for this report tracks 261 total T&S buyers, roughly two-thirds (68%) of estimated spending comes from the largest buyers in the Tech Giant segment.

   However, top-heavy market concentration is a challenge for T&S software vendors because of the existing internal capabilities of large, well-capitalized firms.

   T&S solutions **M&A activity** has also picked up over the past 5 years, and bigger platforms or vendors have acquired as many as ⅔ of all T&S software startups in the past decade⁷. This time period was defined by an increasing market need for T&S talent, both on the vendor side and by platforms, cash on hand for companies to acquire startups, and a point of technological development before the risks of genAI were clearly articulated by the public. While this activity may decrease the number of vendors available on the market for smaller

---

⁷ Companies like Sentropy, which raised $13M from Initialized Capital and more, were acquired by Discord in July 2021, just a year after they launched. In October 2022, Spotify acquired SaaS moderation company Kinzen after they raised ~$3M the year prior. ActiveFence, the SaaS moderation platform with $108M+ in funding, acquired Spectrum Labs in September 2023 after they raised $32M from Intel and Greycroft.
platforms and consolidate software spending with large vendors, T&S software startups are primed for growth today with more acknowledgment from platform customers.

2. **Software efficiency**
   History has shown that technology is deflationary – successful development of T&S technology will allow T&S teams to keep more users safe at a lower cost. Therefore, future gains in market value for software will likely come at the expense of some services spending.

3. Additionally, **BPO employee unionization** is another key consideration that platforms are managing, as content moderators worldwide are raising alarms about their poor working conditions and exposure to traumatizing content. Accelerating in 2023, these workers have been unionizing in regions ranging from Africa to the EU. As partnerships between BPOs and platforms become less predictable, enterprises may be more willing to try new solutions that can fulfill their global needs, such as LLMs that leverage Massively Multilingual Speech (MMS) models to extend moderation across non-dominant languages.

**T&S Solutions 5-Year Forecast**
The number of T&S software companies has been increasing in recent years, which coincides with the increased market demand for these solutions. Within the past 13 years (2010-2023), the number of T&S software companies on the market increased from 73 to 174, a 138% increase.

![Number of T&S Software Companies by Year Founded](image)

Looking ahead, experts say T&S software solutions will rapidly replace a significant portion of T&S services solutions over the next 5 years due to advancements in AI, big data analytics, and the emergence of standard content moderation guidelines. Founders with T&S expertise are building towards an inflection point where automated solutions will become more precise than manual reviewers and costs for automated solutions will drop significantly. This trend is underscored by the influx of early-stage capital to T&S software startups, with ~32% of 2023 funding into Series A, seed, pre-seed startups or accelerator graduates.
Research shows the TAM for T&S software will grow at a CAGR of 15.7% to reach $15.4B in 2028. Furthermore, the SAM for T&S software could grow at a CAGR of 36.6–57.2% depending on how effectively vendors capture demand for software solutions.

However, experts were split on vendors’ ability to capture this increase in spending. One school of thought posits that companies will opt to in-house the majority of the increased software spending (In-house-led). These experts predict that sensitive data, complex integrations, rapid update requirements, and strong in-house development teams will prevent vendors from breaking into the largest potential accounts. In this scenario, the SAM for T&S software grows at a CAGR of 36.6% to reach $3.1B in 2028. This assumes that the Mid-Market and Enterprise segments account for ~$2.5B (70%) of software SAM growth, while the Tech Giant segment accounts for ~$900M (29%) of software SAM growth.

The opposing school of thought posits that vendors will capture much of this spending by providing significant value via proprietary datasets, model training, and support services (Vendor-led). In this scenario, the SAM for T&S software grows at a CAGR of 57.2% to reach $6.3B in 2028. This assumes that the Mid-Market and Enterprise segments account for ~$3.8B (60%) of software SAM growth, while the Tech Giant segment accounts for ~$2.4B (38%) of software SAM growth.
WHY INVEST IN T&S?

A. CAPITAL INVESTMENTS AND EXPERTISE
B. COMPETITIVE LANDSCAPE AND MARKET PLAYERS
C. ROLE OF EMERGING TECHNOLOGY AND OPPORTUNITIES IN T&S
D. CORPORATE BENEFITS OF T&S
A. Capital Investments and Expertise

Venture investors have been funding startups in the T&S solutions market for over a decade. However, the past few years represent an inflection point: increasingly urgent company needs related to T&S, scalable and useful new products, and founders who have the expertise to stay ahead of the market have created a sizable and growing investment opportunity capable of venture scale returns. In addition, the evolution of T&S vendors has allowed enterprises to outsource T&S concern, driving revenue and protecting the brand.

According to Pitchbook, from 2019 to 2023, the total amount of venture capital invested into the T&S software market alone, at $7.78B, was 4.7x the amount of capital invested from the previous 5 years (2014–2018), and 40x the amount of capital invested from 2009–2013. Top firms have taken notice and are actively participating in this ecosystem. Funds like 8VC, Accel, a16z, Amplify Partners, Benchmark, Bloomberg Beta, Union Square Ventures, BoxGroup, Norwest, General Catalyst, and Thrive Capital have invested in the category. At the earliest stages, Y Combinator has conducted 18 deals in 11 T&S software startups since 2011. And emerging funds, such as J2 Ventures and ex/ante, have investment theses directly aligned with supporting this industry.

T&S software median deal sizes were also increasing over the past 5 years, with 2023 having a median deal of $5.26M, up 3x since 2019. These larger deal sizes are characterized by increasing corporate M&A and private equity deals in 2021–2023, demonstrating maturity in the market.
In addition, investment can be broken down by software type. Forty-one percent of all T&S software investment ($1.6B) was dedicated to account/ID verification tools, 32% ($1.3B) was dedicated to OSINT/risk and threat intelligence tools, and 12% ($0.5B) was dedicated to SaaS moderation platforms, according to Pitchbook. Account/ID verification and OSINT/risk and threat intelligence tools have received more investments as these companies are relatively older and have been serving more mature industries, such as cybersecurity. However, the other segments are poised for growth as their offerings mature and overall demand for T&S software increases.

**T&S Software Total Funding Amounts by Category Segment**

![Diagram showing funding amounts by category segment](image)

**New Wave of Investors With T&S Expertise**

The nascency of the T&S sector has created a situation where not every investor understands the T&S market and its opportunities. However, there are a growing number of investors who have dealt with T&S issues firsthand as operators (see Appendix H - Expert Investors) and can understand the differences in how different teams building in the space develop solutions to solve critical issues that platforms have historically addressed ad hoc, manually, and/or by leveraging solutions that are not purpose-built for T&S. Thus, the runaway success of platforms over the past 10+ years has not only yielded a talent pool of founders capable of building the next generation of T&S startups, but also investors who can spot new opportunities in the space and the right teams that are capable of exploiting them.

**Industry Drive to Protect National Security**
T&S is a key part of the national security conversation. For example, in 2023, RAND researchers published a paper calling “truth decay”, or the “diminishing role of facts and analysis in public life”, a U.S. national security risk describing it as a “huge vulnerability” and a “strong weapon” in the hands of adversaries.

At an industry level, there is a renewed focus on investing in companies that protect the national interest. Building on this momentum, in 2023 VCs allocated $7B to defense tech startups. While advanced AI has escalated these issues on platforms, T&S solutions can help companies fortify themselves against these types of risks as well as foreign interference.

**Global Investment Outlook**

T&S software investments are global in nature. Investors based in the U.S. deployed 64% of total capital, European investors deployed 19% of total capital, while investors based in other global regions invested 16% of total capital. While the U.S. has the most active portfolio in this industry, non-U.S. funds such as Enterprise Ireland, Dawn Capital, Adara Ventures, and AlbionVC are investing in the growth of T&S software companies outside of the United States.

**B. Competitive Landscape and Market Players**

Today’s global competitive landscape for T&S solutions is growing to meet the demands of a quickly evolving industry. And investment in T&S software has been robust. In the SaaS moderation platform segment, Hive, founded in 2013, has raised $165M from General Catalyst, 8VC, Glynn Capital, and Visa Ventures among others, and ActiveFence, founded in 2018, has raised $108.8M from CRV, Grove Ventures and Norwest. In the specialized tools segment, Siren, founded in 2014, has raised $26.7M from Atlantic Bridge Capital, Frontline Ventures, and Growing Capital. Investors are active in all of these spaces and there are many more examples of successful investments (see Appendix C – T&S Vendor Overview).
C. Role of Emerging Technology and Opportunities in T&S

Artificial Intelligence and T&S
The advent of generative AI marks a transformative era in the T&S industry. In the age of AI, platforms expect to be managing increasing volumes of AI-generated content across their applications. In anticipation of this, dedicated providers and leading technology companies are developing solutions to manage this content also using generative AI technologies.

The combination of spikes in synthetic content, increasingly reliable generative AI tools, and firms’ eagerness to integrate those tools create opportunities for commercial T&S offerings. For example, Unitary, which raised ~$25M, is building “context-aware AI” to improve detection and classification of visual content. Their application of multimodal AI models has made it possible for platforms to moderate video more effectively than with “visual only models” that have been the industry standard for years. Additionally AssemblyAI, which secured ~$108M in funding, offers speech recognition models that aid in audio content moderation. As compute costs recede, new and established players will continue to seize upon opportunities to offer T&S services.

The Future of T&S
While many T&S solutions aim to accurately apply policy-directed rules across various types of content, AI models have demonstrated an increasing skill in parsing the meaning of complex rules and applying them to nuanced content. T&S leaders and investors are bullish on this emergent capacity as a key supplement to traditional content moderation, and even as a replacement in the long term. Effective integration of new AI tools may also reduce the need for human review of the most harmful content, mitigating psychological damage on many thousands of content moderators.

8 At the 2024 Munich Security Conference, 20 leading technology companies, including OpenAI, X, Amazon, and Google, pledged to work together on addressing harmful AI content.
In 2023, Mayfield Fund published “Trust and Safety Cannot Be An Afterthought”, highlighting their commitment to supporting founders who build companies that address known harms and plan for unforeseen risks. While their definition of T&S goes beyond just the T&S solutions market, the overlap is apparent and indicative of how the sector may evolve in the coming years. The market demand for T&S software in the age of AI is anticipated to grow at a CAGR of 36.6%–57.2% over the next 5 years as risks become more apparent and acute.

Opportunities in Content Moderation
GenAI tools may be combined with existing systems in ways that increase efficiency. For example, LLMs can analyze and A/B test new content policy proposals, or even suggest new ones in a collapsed timetable. Additionally, some believe genAI tools will enable more forms of community governance, reducing dependence on centralized T&S policy enforcement teams.

Vendors are also capitalizing on genAI capabilities to augment their services. Available LLMs now provide a strong baseline from which they may customize their models to produce acceptable results at a lower expense, enabling a bigger shift from BPO services. As these LLM-enabled software services mature and become more refined, there will be greater opportunities to replace expensive BPO contracts with integrated tool-based services.9

Other Emerging T&S Opportunities
The $249B gaming industry also has emerging opportunities for T&S solutions. Even though T&S challenges across gaming platforms have attracted less attention than social media firms, gaming providers have been making heavy investments into player safety tools9 like Modulate, especially as many provide multiplayer and extended reality (XR) experiences. AI-enabled T&S tools will be particularly valuable to both established and new gaming firms, and are a sizable business opportunity for T&S vendors.

Additionally, as 4.1 billion people enter an election cycle in 2024, experts anticipate platforms will seek additional support to manage the heightened risks of election interference and mis/disinformation campaigns. Acute military conflicts such as the Israel/Gaza and Russia/Ukraine wars will bring increased focus on managing violent content, information integrity, and persuasion campaigns. Both public pressure and regulatory interest in these areas are likely to increase the need for demonstrable investments in emerging T&S software that manages these threats, such as watermarking and information provenance software.

And as a growing trend of youth (97% in the U.S.) access the internet daily, there is an increasing need and requirement for robust age verification technologies. Some have suggested that LLMs

9 BPOs are also noticing this trend – for example, BPO TaskUs invested ~$2M in AI-based content detection company Light in Nov 2022 to enhance their AI-enabled capabilities.
10 Microsoft acquired SaaS moderation platform Two Hat in Oct 2021 to implement proactive moderation technologies for its Xbox, Minecraft, and MSN gaming communities.
could improve age verification measures by analyzing a user’s posts or messages to flag content likely created by someone who is underage.

**Challenges to Adoption**

Firms considering LLM and other generative AI integrations must navigate certain challenges. The efficacy and speed of compute-heavy LLMs can make genAI solutions cost prohibitive, especially at scale. Institutional inertia may favor reliance on legacy content classifiers, which are often integrated across databases. Additionally, explainability of closed-source model decisions may be a challenge, especially as platforms are expected to improve their transparency efforts and gain user trust.

However, as compute costs continue to decline and these tools are refined according to their customers’ needs, it is likely that genAI will be welcomed across T&S efforts.

**D. Corporate Benefits of T&S**

While T&S solutions are ripe for investment from venture capitalists, making investments at the corporate level is also critical to ensuring business success. And, for some firms, that means investing in an outside vendor for T&S solutions.

A few predictable forces can influence T&S budget. In addition to the business pressures previously identified as market drivers, companies that have products with significant “real world” risks are estimated to have higher T&S budgets compared to those that don’t. For example, if a user listens to pirated music on a streaming platform, the stakes are fairly low. However, if a user goes on a date with a predator from a dating site, there is potential for serious harm. Other company characteristics that affect T&S budget include level of usership, number of different product types, overall company revenue, number of content types, and overall company headcount (see Appendix D – Drivers of Trust & Safety Budget).

While companies should use these considerations to ultimately determine when it’s most appropriate to engage a vendor, when deployed, T&S vendors can greatly benefit business platform operations in terms of efficiency, flexibility, and efficacy. This can reduce the time it takes to review and take action on issues, prevent overburdening internal staff, and enhance the efficacy of content moderation.

1.  **Efficiency**: Internal teams can optimize their time focusing on core, strategic issues as vendor solutions reduce the time needed to make key decisions. Brian Fishman, chief strategy officer at Cinder, has noted that since their clients have reduced active handle time by 90+%, clients are now able to use the time saved to better address backlogs or previously unmoderated areas. Additionally, vendors reduce the time needed to ramp up new workflows from weeks or months to hours or days, enabling T&S teams to not only address their current T&S issues, but also experiment towards improved approaches.

---

11 To account for this risk, the Report incorporated an “IRL Risk Premium” of 30% to the reference budget for companies with such products for the market sizing analysis. See Appendix B – Methodology for more details.
2. **Flexibility:** Vendors provide access to a diverse pool of efforts and expertise with more flexibility, whereas internal capabilities may be limited by subject matter or time constraints. The availability of vendors also helps avoid the lengthy process of hiring full-time internal team members.

3. **Efficacy:** Vendor solutions incorporate lessons from a broad landscape of T&S risks observed across their numerous platform clients, enabling platforms to employ more proactive and accurate methods for addressing T&S risks. Additionally, vendor tools can protect internal team members from irrelevant or troubling details, which enable them to focus on the critical tasks at hand.

Though it can be difficult to measure the direct, positive effects of T&S efforts due to the nature of T&S (if all is working well, there will be no observable issues to report), there are a few predictable patterns whereby T&S investments foster positive outcomes:

1. **User engagement, retention, and growth**
   As user’s exposure to problematic content increases, they’re more likely to leave the platform. But, when users feel secure, they are more likely to engage in meaningful interactions and produce more organic content, which contributes to a healthier online community and supports overall user growth (see Appendix G - Case Studies: Minecraft).

Matthew Soeth, former head of community at Spectrum Labs, has noted that their clients have seen an increase of customer LTV by 30% by reducing harmful content and promoting healthy behaviors.

**Case Study 1: Minecraft**

Minecraft is a gaming platform with 174M monthly users, 63% of whom are 21 and under. Microsoft acquired its gaming provider, Mojang, for $2.5B in 2014. With more than 300M unique copies sold as of late 2023, Minecraft is the best-selling game of all time.

After 120k parents protested the platform in 2020 due to child predation concerns, Mojang has since improved parental controls, player safety education, reporting tools, among other key T&S improvements. In October 2021, Microsoft acquired Two Hat to support its safety efforts in its growing gaming portfolio.

To date, child safety media reviewers have endorsed Minecraft, with Common Sense Media rating it four stars for children ages 8+ and Pan-European Game Information giving an age rating of 7+, supporting its popularity among parents and their children.

2. **Revenue and funding**
   Increased user engagement from T&S investments will fuel user-based revenue streams such as subscription revenue and on-platform transactions (see Appendix G - Case Studies: Airbnb), while users with poor experiences will spend less money. According to ADL’s Hate and Harassment in Online Games 2023 survey, 20% of adults and teens are spending less money within online games due to the hate and harassment they face on these platforms.

---

12 Jeff Horwitz’s book about Facebook, Broken Code, reveals that “integrity work has boosted usage among new users by 0.9%” and overall usage by “a fraction of a percent”.

---

**TRUST & SAFETY MARKET RESEARCH REPORT** by Duco
Case Study 2: Airbnb

Airbnb has pioneered peer-to-peer accommodations in the sharing economy, and was valued at $86.5B upon its IPO in December 2020 – more than the combined worth of the 3 largest hotel chains. While the company was host to numerous public crises and reactive safety commitments in its early years, within the past 5 years, Airbnb has demonstrated a more publicly proactive T&S investment strategy.

Airbnb has an ethos of proactively designing for trust, reflected in its platform features, formation of a T&S Advisory Coalition with 22 organizations, and acquisition of T&S SaaS moderation platform Koko, according to interviews with former employees. Their most profitable quarter ever due to increased revenue streams from hosts and guests was enabled by these investments into positive customer experiences and essential T&S offerings. As it continues to grow in market dominance relative to hospitality incumbents, experts estimate that Airbnb’s T&S solutions budget will grow approximately $5.8M at a CAGR of 9% in the next 5 years.

Furthermore, brands will subsequently take notice of platforms with conditions for healthy ad revenue and spend advertising dollars accordingly. However, when platforms pair ads with unsafe content, advertisers may suspend ad spend to protect their brand (see Appendix G - Case Studies: X (formerly Twitter)), especially when there is a lack of transparency over which content categories they advertise on.

Case Study 3: X (formerly Twitter)

Since Elon Musk’s acquisition of X (formerly Twitter) and divestment of T&S solutions that have led to a rise in hate speech on the platform, big advertisers have removed their spending – leading to a potential loss of up to $75M in revenue.

To T&S professionals, even among Twitter’s past crises, this marked a major shift from the company’s previous proactive investments into T&S, such as Rethink, a prompt that reduced potentially offending tweets, and its 2018 acquisition of T&S moderation platform Smyte. These investments paralleled the platform’s user growth from 2017–2022, when monetizable daily active users grew 118%. However, under Musk’s leadership and minimalist T&S approach, X has lost approximately 13% of daily users from 2022–2023.

And for platforms fundraising, opportunities may be influenced by the strength of a company’s T&S programs, as investors believe that this is an indicator of longevity. Michael Dworsky, CEO of Cove, has been approached by multiple companies pre-launch due to investor pressure for a T&S solution early in the platform life cycle.

In addition, investing in T&S will enable platforms to generate cost savings associated with efficiency gains, resilient employees, preventing regulatory costs, interruptions to focus, and brand safety issues. A strong T&S program will also help the company mitigate the risk of harm as well as address systemic platform risks.

---

13 Content governance practices have had increasing investor interest on earning calls in recent years (see p.11 for examples).
1. **Cost savings**
Efficiency gains from vendor support, such as ActiveFence’s ActiveOS users experiencing a **38% improvement in operational efficiency**, can create cost savings. Additionally, a T&S program that proactively detects and remedies issues is less likely to face the growing body of **regulatory and legal investigations, fines, and potential suspensions**.

2. **Risk reduction**
Vendors that provide comprehensive detection, scaled enforcement capabilities, and knowledge help platforms stay ahead of evolving risks before they cause real-world harm to their users. They also help platforms better anticipate and recognize T&S risks at the system-level.

For more details on how T&S vendors can enhance business operations, improve profitability, and reduce costs, see **Appendix F – Corporate Benefits of T&S Vendor Partnerships**.
CONCLUSION

A. THE TIME FOR INVESTMENT

B. OPPORTUNITIES FOR ADDITIONAL ENGAGEMENT
A. The Time for Investment

Managing content- and conduct-related risks, catalyzed by the recent proliferation of advanced AI systems, is one of the most challenging problems facing the technology industry today.

Regulators around the world are passing laws to make technology companies liable and require that they manage these risks. And founders with deep expertise in T&S are starting companies to meet platforms’ growing demand for effective and scalable solutions that help lessen subjective decision making and enable them to build products that better reflect the needs of their users. According to the Atlantic Council, “a 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.”

While T&S risks have historically been addressed through an intricate combination of human and machine intelligence, recent advancements in AI systems have facilitated an unbundling of the T&S software stack, potentially drawing market share away from manual services like BPOs and creating new investment opportunities in a rich ecosystem of AI-enabled T&S software startups that are poised to grow as harms become more acute in the coming years.

Those potential harms – and the companies responsible for mitigating them – are under a microscope. In January 2024, the U.S. Senate held a hearing about online child safety where elected officials questioned the CEOs of TikTok, Discord, X, Meta, and Snap about their practices to protect users. Twenty years after Meta was founded, executives are still subjected to inquiries about how they are managing known risks. And with the onset of advanced AI systems, the risk landscape is poised to broaden and intensify. The type of transparency that T&S solutions afford gives regulators, users, and broader civil society access to critical and necessary information.

The goal of this report is to bring awareness to a burgeoning category of new and established T&S solutions that address acute challenges for all platforms hosting user-generated content and/or services. The sector has quickly evolved from a vitamin to a painkiller, driven by emerging regulations and new compliance requirements globally. Fostering an ecosystem of T&S solutions will facilitate knowledge sharing among founders, operators, and enterprise customers, particularly as new risks emerge.

Venture capitalists have been funding startups that fall under the T&S solutions category for many years, totaling billions of dollars in capital. However, recognizing T&S as a distinct industry is critical in order to fuel more investment into related companies and organizations, professionalize the field, enable competitive vendor benchmarking, support policymakers in developing right-sized regulations and industry recommendations, and encourage necessary multisectoral collaboration. The market is growing and critical to the advancement of society and, importantly, it is ripe for investment.
Companies and platforms must also pay attention to the growing risks associated with user and AI-generated content and make appropriate investments in T&S, especially to identify and track key metrics that make a clearer business case for why managing these risks should be operating norms. Beyond typical outsourcing considerations like vendor cost, suitability of their solutions, and diversity of providers when evaluating potential vendors, it’s helpful to evaluate a few attributes that will enable a successful T&S vendor partnership. First, T&S vendors should be collaborative. Every company has unique T&S needs, systems, and people; vendors should be able to flex their capabilities to align with companies’ strategic goals, without creating friction or distracting internal T&S teams. Second, T&S vendors should be trustworthy: collaborating with vendors may involve sharing sensitive company information that should not be made public (e.g., existing risks). Finally, it’s helpful to look for vendors who are able to grow together with the T&S program. Once a company finds an effective vendor, it may be beneficial to expand the vendor’s scope to new areas, rather than introducing new vendors who will take time to become an expert in the company’s T&S risks, integrate with its systems, and more.

To support the growth of the T&S industry, platform leaders should also consider sharing their knowledge on T&S solutions and provide greater transparency on the tools they use to manage their T&S issues. This would help younger platforms understand how to navigate the T&S market and what T&S solutions they should consider as they mature.

B. Opportunities for Additional T&S Engagement

This Report’s outline of the $14.7B SAM for T&S solutions, its growing investment potential, and corporate benefits indicate that there are ripe opportunities for further growth and engagement with the T&S industry. Below are additional recommendations for various stakeholders to support this growth:

Philanthropic and Public Funding

1. Continue the momentum of the Trust & Safety Hackathon to identify and build innovative solutions that address T&S challenges, especially those that persist outside of the Western Hemisphere.
2. Develop a non-profit open source T&S Tooling Accelerator to enhance accessibility of T&S tools for emerging platforms. Partner with platforms willing to engage in corporate philanthropy for supplemental resources and talent.
3. Fund accelerators to embed T&S solutions in the beginning of early-stage platform development.
4. Develop networks and partnerships with aligned funding institutions across the public and private sectors to support the growth of this industry.
5. Continue building coalitions and transparency initiatives across academia, industry, government, and civil society.
T&S Solutions Vendors

6. Offer lower-cost versions of solutions for emerging platforms, which can build case studies for future sales.
7. Leverage AI capabilities to provide stronger and more cost effective product offerings.
8. Publish best practices, case studies and other thought leadership to both build brand eminence and share critical knowledge across the industry.

Professional Organizations (e.g. TSF, TSPA, Integrity Institute)

9. Build upon the T&S solutions stack to identify vendors that provide these services to help platform leaders understand which solutions are available to them.
10. Establish a global classification system of digital harms and transparency measures based upon product and content types to harmonize cross-industry standards and best practices among platforms, T&S vendors, and investors.
11. Develop best practice performance metrics to quantify the impact of T&S solutions.
12. Develop professional T&S curriculum (e.g. business case studies, courses, certifications) to train current and aspiring investors and platform leaders.

Academia and Civil Society

13. Continue to conduct and publish research on emerging T&S challenges and best practices, including those related to adopting T&S solutions themselves.
15. Partner with policymakers to explore legislation that supports public investment in T&S solutions.

Government Bodies

16. Collaborate with T&S professional organizations to craft informed and technically feasible legislation on content and transparency guidelines, safe development of open source AI models, and data-sharing across institutions.
17. Require platforms to provide data to apps which enable users to block hateful speech (e.g. Block Party, Apollo), boosting the D2C T&S ecosystem.
Appendix A

Glossary of Key Terms

- **Account/ID verification**: Products that verify user accounts or identities, including age verification. This is 1 of the 4 domains of T&S Specialized Tools.
- **Business process outsourcing services (BPOs)**: Outsourcing firms that companies use to help with content moderation and managing T&S risks. While T&S BPOs provide a spectrum of services, such as risk advisory, policy development, customer service management, or enterprise modernization, content moderation or “content trust and safety” comprises a sizable part of the T&S BPO market.
- **Content moderation**: The process of reviewing online user-generated content for compliance against a digital platform’s policies (community guidelines) regarding what is and what is not allowed to be shared on their platforms (TSPA). Moderation is a specialized process that often requires human intervention to review and filter content for compliance.
- **Community guidelines**: The set of conditions and limitations governing use of a digital service that a user must agree to as a condition of use. (DTSP)
- **Detection models/APIs**: Includes 3rd party models and/or APIs that platforms use to detect and/or take action on problematic content or conduct. This is 1 of the 4 domains of T&S Specialized Tools.
- **Digitization**: The integration of digital platforms into everyday life.
- **OSINT/risk & threat intelligence**: Open source intelligence, risk intelligence, and threat intelligence tools collect, process and analyze data to understand emerging hazards or trends. This is 1 of the 4 domains of T&S Specialized Tools.
- **SaaS moderation platforms**: Today’s T&S software-as-a-service (SaaS) moderation platforms power various parts of the T&S workflow, which includes creating policies, detection, taking action, measuring progress and/or sharing results. This is purpose-built software to specifically manage the complex operational issue of user-generated content.
- **Serviceable addressable market (SAM)**: The T&S solutions Serviceable Addressable Market (“SAM”) represents the portion of the T&S solutions TAM that companies are currently outsourcing today.
- **Specialized tools**: Tools designed to provide insights that aid T&S teams in their efforts from one of the following domains - OSINT/Risk & Threat Intelligence, Account/ID Verification, Watermarking/Information Provenance, Detection Models/APIs. While these technologies may be used by a variety of functions outside the realm of T&S responsibilities (e.g. cybersecurity, fraud management), they are included in the analysis since these capabilities serve a critical role providing expertise among various aspects of the T&S workflow.
- **T&S buyer**: Digital platforms that host user-generated content or services and are seeking solutions to manage these content or services according to their community guidelines. These platforms include social media, search engines, messaging applications, streaming services, gaming, dating apps, the sharing economy, web services, GenAI providers, and app stores.
- **T&S solutions market**: The diverse marketplace designed to help T&S teams with everything from policy development to detecting allowable content and scaled moderation operations.

- **Total addressable market (TAM)**: This report features a bottom-up analysis of the TAM for T&S solutions. Our analysis counts the total number of companies that could employ T&S solutions, and estimates each company’s annual spend.

- **Trust & safety (T&S)**: The policies, processes, and tools digital platforms employ to manage content- and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety (DTSP)

- **User-generated content (UGC)**: UGC includes media, posts, or comments shared by users on a platform (TSPA)

- **User-generated services (UGS)**: UGS includes services provided by users on a platform (e.g. lodging, driving, delivery)

- **Watermarking/information provenance**: Products that identify the origins of media online, used to help users distinguish between human, AI-generated, and/or edited content across audio, image, text or audio. This is 1 of the 4 domains of T&S Specialized Tools.
Appendix B

Methodology

Below is an overview of the Report’s methodology for analyzing both the solutions and buyers in the T&S Market:

I. T&S Solutions Analysis

To understand what T&S solutions exist on the market, researchers used Harmonic, a startup database that indexes 20M+ companies based on website information and filings and identified 222 private sector companies that fall under the T&S solutions scope. The Report is focused on T&S companies founded in 2010 or later, leveraging sample search terms including “trust & safety”, “content moderation”, “misinformation”, and/or “risk intelligence” and conducted a bottom-up analysis to segment core offerings. Firms identified have a clear presence in the market, demonstrate an active provision of commercial activity related to T&S tooling, and have identifiable revenue, employment or trading activity. Researchers then cross-referenced the Harmonic dataset with Pitchbook for investment analysis.

While many of the Tech Giant companies are also involved in the development of T&S solutions (e.g. AWS’ Rekognition Image Moderation API), they have been excluded from the analysis since this report is focused on dedicated third-party T&S providers. The Report also excludes firms no longer under operation.

II. T&S Buyer Analysis

To understand potential T&S buyers on the market, researchers identified 261 companies that could consume T&S solutions based on their core offerings. They then queried business databases (LinkedIn Sales Navigator, Apollo.io, and Harmonic.ai) to create this list of companies with T&S employees on staff. And then searched for those companies’ competitors to round out the hypothetical T&S solutions customers list.

Once the list was complete, company information was added. Data points collected included: estimated/announced active user metrics (DAU, MAU, others), products offered (marketplace, social media, streaming, etc), content types used by those products (audio, video, live streaming, various types of text content), estimated revenue, employee counts, public trading status, and special factors like in real life interaction risk, and children content offered.

Guided by expert interviews with current and former T&S professionals, the list was segmented into three size buckets of customers:

---

14 This is consistent with previous adjacent studies, such as Paladin’s International State of Safety Tech 2023 Report.
Experts also estimated “reference T&S solutions budgets” for each size segment. These reference budgets represent the hypothetical maximum a company in that size segment could responsibly spend on T&S software and services. To complete the bottom-up market sizing estimates, researchers scored each company based on the data points collected to estimate how close the company’s actual budget might be to its respective reference budget. The sum of these budget estimates makes up the Total Addressable Market (TAM). See this spreadsheet for additional analysis on estimated T&S budgets and an interactive TAM/SAM forecast that adjusts based on different segment reference budgets, factor weights, and other variables.

All analysis is up to date as of January 2024.

Limitations

T&S Solutions Analysis: Data for this report was sourced from Harmonic, which indexes millions of startups and provides metadata on key company attributes. BPOs are not necessarily considered startups and so may be undercounted in the report, although ~20% of the sample does fall into this category. Additionally, since T&S is a nascent industry, startups may not be using terminologies included in the keyword search list. However, this dataset accounts for the majority of T&S solutions globally.

T&S Buyer Analysis: Market sizing data for this report was sourced from LinkedIn Sales Navigator and Apollo.io. While there are likely companies that could consume T&S which are not tracked by this report, experts believe this dataset accounts for the majority of T&S spending globally.

Additionally, the “bottom-up” market sizing approach requires an estimate of the spending potential of consumers in the market. The Report used externally observable data points and expert interviews to estimate T&S budgets across customer segments. Therefore, financial figures in this report should be viewed as estimates.
Appendix C

T&S Vendor Overview

T&S Solutions Stack

The chart below demonstrates how different types of solutions can be leveraged in concert or separately to create a T&S solutions stack that fulfills the T&S needs of technology platforms. For example, a platform may employ a BPO for activities related to detection, taking action, or measuring progress. Progress is then shared with internal platform teams, which can package results and share them with relevant stakeholders, like the Board of Directors, shareholders on earnings calls, journalists, etc.

![Trust & Safety Solutions Stack](image)

1. T&S Business Process Outsourcing Services

Business processing outsourcing services (BPOs)\(^\text{15}\) are outsourcing firms that companies use to help with content moderation and managing T&S risks. While T&S BPOs provide a spectrum of services, such as risk advisory, policy development, customer service management, or enterprise

\(^{15}\) Data for this report was primarily pulled from Harmonic, which indexes millions of startups and provides metadata on key company attributes. BPOs are not necessarily considered startups and may be undercounted in the report, although ~20% of the sample does fall into this category.
modernization, content moderation or "content trust and safety" comprises a sizable part of the T&S BPO market.

BPO employees who are subcontracted for content moderation services fulfill a variety of functions, including but not limited to:

- Assess potentially harmful content for adherence to company guidelines and legal/reporting requirements
- Tag content with additional metadata to help train AI systems to detect content automatically
- Flag problematic content for further review
- Triage content to other teams
- Take action on violating content

Here are some sample companies with descriptions pulled directly from the company website:

**Atento**

Atento is the largest provider of customer relationship management (CRM) and business process outsourcing (BPO) services in Latin America, and among the top five providers globally, based on revenues.

<table>
<thead>
<tr>
<th>Founded</th>
<th>1999, Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>93,285</td>
</tr>
<tr>
<td>Latest round</td>
<td>10/2014, IPO (5/2020, PIPE)</td>
</tr>
</tbody>
</table>

**TaskUs**

TaskUs is a collective of highly capable humans, who understand how to deploy technology and data to best serve their clients. From Digital CX to Trust & Safety, AI Services, Risk + Response, Consulting, the company works to protect their clients’ interests and support their long term success through innovation and technology.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2018, Texas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>25,089</td>
</tr>
<tr>
<td>Latest round</td>
<td>6/2021, IPO (7/2022, PIPE - II)</td>
</tr>
<tr>
<td>Raised</td>
<td>$279M from Blackstone; Navegar; Western Alliance Bank</td>
</tr>
</tbody>
</table>
Teleperformance is a global digital business services company that offers services for content moderation. The company’s worldwide reach and regional presence enable it to utilize its international staff for addressing local issues. The firm delivers digitally-powered business services to help the world’s best brands streamline their business.

<table>
<thead>
<tr>
<th>Founded</th>
<th>1978, France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>215, 480</td>
</tr>
<tr>
<td>Latest round</td>
<td>01/2007, IPO</td>
</tr>
</tbody>
</table>

**TELUS International** designs, builds and delivers next-generation digital solutions to enhance the customer experience (CX) for global and disruptive brands. TELUS International's integrated solutions and capabilities span digital strategy, innovation, consulting and design, digital transformation and IT lifecycle solutions, data annotation and intelligent automation, and omnichannel CX solutions that include content moderation, trust and safety solutions, and other managed solutions. TELUS International partners with brands across high growth industry verticals, including tech and games, communications and media, eCommerce and fintech, healthcare, and travel and hospitality.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2005, Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>31573</td>
</tr>
<tr>
<td>Latest round</td>
<td>02/2021, IPO</td>
</tr>
</tbody>
</table>

2. **T&S Software**
   
a. **SaaS Moderation Platforms**

Today’s T&S SaaS moderation platforms power various parts of the T&S workflow. Prior to these tools, platforms leveraged ticketing systems such as Zendesk designed for customer service use cases to track and moderate billions of content issues. T&S SaaS moderation vendors have since created purpose-built software to specifically manage the complex operational issue of user-generated content for T&S teams. Technological advancements have enabled these vendors to incorporate best in class third party AI detection models, APIs, risk intel features and other customized solutions for their customer platforms’ specific needs.

Here are some sample companies with descriptions pulled directly from the company website:
**ActiveFence** is a T&S provider for online platforms, protecting platforms and their users from malicious behavior and content. ActiveFence has been expanding its product offerings through the acquisition of T&S startups including Spectrum Labs and Rewire, both in 2023. The company currently serves clients including Cohere, Stability AI, Deliveroo, and Grindr.

<table>
<thead>
<tr>
<th><strong>Founded</strong></th>
<th>2018, Israel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current headcount</strong></td>
<td>290</td>
</tr>
<tr>
<td><strong>Latest round</strong></td>
<td>July 2021, Series B ($100M)</td>
</tr>
<tr>
<td><strong>Raised</strong></td>
<td>$108.8M from CRV, Grove Ventures, Norwest Venture Partners</td>
</tr>
</tbody>
</table>

**Cinder** is the industry’s first T&S operations platform to help organizations combat Internet abuse at scale. This tech startup provides T&S teams with a single system to manage complex integrity operations and investigations to create a safe environment for their users. The co-founding team has extensive T&S experience at Meta, Palantir, and the U.S. government.

<table>
<thead>
<tr>
<th><strong>Founded</strong></th>
<th>2021, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current headcount</strong></td>
<td>26</td>
</tr>
<tr>
<td><strong>Latest round</strong></td>
<td>December 2022, Series A ($12.8M)</td>
</tr>
<tr>
<td><strong>Raised</strong></td>
<td>$14M from Accel, YC, J2 Ventures</td>
</tr>
</tbody>
</table>
**Cove** provides a no-code platform that enables customers to remove abusive content and behavior from their products at scale. The startup leverages cutting-edge AI and makes it accessible through powerful abstractions and UIs so our customers can move quickly and thoughtfully. Cove’s product supports policy development, policy enforcement, and user report review workflows. Currently, Cove supports companies including Yik Yak and Sidechat.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2021, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>10</td>
</tr>
<tr>
<td>Latest round</td>
<td>July 2023, Seed ($5.8M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$5.93M from Common Sense Growth, Thrive Capital, Sound Ventures, Neo</td>
</tr>
</tbody>
</table>

**Hive** specializes in cloud-based AI solutions for content analysis. Its services include AI-driven sponsorship measurement, marketplace authenticity, age verification, language translation, and content moderation. The company caters to various sectors, including automotive, financial services, media, communications, manufacturing, travel, retail, and restaurants.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2017, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>401</td>
</tr>
<tr>
<td>Latest round</td>
<td>April 2021, Series D ($50M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$165M from General Catalyst, 8VC, Glynn Capital, Visa Ventures, Daniel Rumenik, Vic Singh, Maiden Lane</td>
</tr>
</tbody>
</table>

**b. Specialized Tools**
T&S specialized tools are designed to provide insights that aid T&S teams in their efforts from one of the following domains:

1. **OSINT/Risk/Threat Intelligence**
   
   Open source intelligence, risk intelligence, and threat intelligence tools collect, process, and analyze data to understand emerging hazards or trends.

Here are some sample companies with descriptions pulled directly from the company website:

**BLACKBIRD.AI**

Blackbird.AI, a global leader in Narrative & Risk intelligence, delivers advanced AI-powered solutions in its mission to empower trust, safety and integrity across the information ecosystem. The platform offers solutions such as a constellation dashboard, risk intelligence engine, risk intelligence ops, media security, and more services.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2017, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>49</td>
</tr>
<tr>
<td>Latest round</td>
<td>June 2023, Series B ($20M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$30.53M from Generation Ventures, Redpoint, Ten Eleven Ventures, Dorilton Ventures</td>
</tr>
</tbody>
</table>
**DarkOwl**

DarkOwl is a Denver-based company that empowers organizations to continually improve their cybersecurity defenses using darknet intelligence. The DarkOwl Vision platform provides access to the world’s largest index of DARKINT™ (darknet, deep web and high-risk surface web) content, along with the tools and services to efficiently find leaked or otherwise compromised sensitive data.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2009, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>41</td>
</tr>
<tr>
<td>Latest round</td>
<td>October 2019, Growth Equity ($8M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$16.64M from Five Elms Capital</td>
</tr>
</tbody>
</table>

**Dataminr**

Dataminr is a leader in real-time event and risk detection within the technology and AI industries. The company provides an AI-powered platform that identifies early indications of significant events and emerging risks from public data, aiding effective crisis response. Dataminr serves a diverse clientele, including global corporations, public sector agencies, and newsrooms.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2009, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>884</td>
</tr>
<tr>
<td>Latest round</td>
<td>March 2021, Series F ($475M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$1.044B from Fidelity, Morgan Stanley, Venrock, IVP</td>
</tr>
</tbody>
</table>
2. Detection Models/APIs

Companies provide AI models and APIs to assist platforms in detecting specific types of content, ranging from audio, to video, and various types of written content.

Here are some sample companies with descriptions pulled directly from the company website:

**Siren**

Siren specializes in investigation data automation. The Siren platform enables users to uncover relationships within multiple data sets, featuring functionalities like search, dashboards, analytics, knowledge graphs, and alerts. Its range of services includes data intelligence, graph analytics, big data search, real-time search, and investigative intelligence.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2014, Ireland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>82</td>
</tr>
<tr>
<td>Latest round</td>
<td>February 2023, Series B ($12.72M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$25M from European Investment Bank, Atlantic Bridge Capital, Frontline Ventures</td>
</tr>
</tbody>
</table>

**AssemblyAI**

AssemblyAI, a developer of AI-driven models, specializes in speech transcription and comprehension. The company offers a product that enables users to automatically transcribe audio, video, and live audio streams into text using a speech-to-text application programming interface (API).

<table>
<thead>
<tr>
<th>Founded</th>
<th>2017, California</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>121</td>
</tr>
<tr>
<td>Latest round</td>
<td>December 2023, Series C ($50M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$108.12M from Y Combinator, Accel, Insight Partners</td>
</tr>
</tbody>
</table>
**3. Account/ID Verification**

Products that verify user accounts or identities, including age verification.

Here are some sample companies with descriptions pulled directly from the company website:

<table>
<thead>
<tr>
<th>Company</th>
<th>Founded</th>
<th>Current headcount</th>
<th>Latest round</th>
<th>Raised</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clarifai</strong></td>
<td>2021, Washington D.C.</td>
<td>98</td>
<td>October 2021, Series C ($60M)</td>
<td>$100.12M from NEA, Manlo Ventures, GV, USV, Lux Capital, NVIDIA</td>
</tr>
<tr>
<td><strong>NuMind</strong></td>
<td>2022, Boston</td>
<td>9</td>
<td>March 2023, Seed ($3M)</td>
<td>$4.5M from Flybridge, Y Combinator, Pioneer Fund</td>
</tr>
</tbody>
</table>
**AUITIX**, an identity intelligence leader headquartered in Israel, provides critical, modular solutions to link physical and digital identities so that companies and their customers can confidently connect.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2010, Israel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>197</td>
</tr>
<tr>
<td>Latest round</td>
<td>November 2019, growth Equity ($20M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$80M from TPG, Oak HC/FT</td>
</tr>
</tbody>
</table>

**Socure** is a digital identity verification platform that verifies identities in real-time and eliminates identity fraud for applicants on the internet. Their predictive analytics platform applies AI and ML techniques with online/offline data to verify every element of identity in real-time, reducing fraud rates and compliance costs.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2012, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>505</td>
</tr>
<tr>
<td>Latest round</td>
<td>November 2021, Series E ($450M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$743.25M from Accel, Initialized Capital, Scale Venture Partners, Generation Ventures</td>
</tr>
</tbody>
</table>
Trulioo is dedicated to verifying the identity of anyone and any business around the world. The company’s identity platform drives global growth for organizations by helping them navigate the challenges of compliance and providing real-time verification of more than 5 billion people and 300 million businesses worldwide.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2011, Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>412</td>
</tr>
<tr>
<td>Latest round</td>
<td>June 2021, Series D ($394M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$469.74M from Goldman Sachs, Blumberg Capital, TCV</td>
</tr>
</tbody>
</table>

Yardstik specializes in offering monitoring and security services to businesses. Their platform is equipped with services including identity verification, background checks, and various screening tools. These features are tailored to assist businesses in efficiently screening, verifying, and hiring individuals on a large scale.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2019, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>50</td>
</tr>
<tr>
<td>Latest round</td>
<td>October 2023, Series A – II ($12M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$24M from MissionOG, Grotech Ventures, Crosslink Capital, Rally Ventures</td>
</tr>
</tbody>
</table>

4. Watermarking/Information Provenance

Products that identify the origins of media online, used to help users distinguish between human, AI-generated, and/or edited content across audio, image, text, or video. The Department of Commerce recently announced that they will be establishing standards for best practice for detecting AI-generated content.

Here are some sample companies with descriptions pulled directly from the company website:
### Pex

Pex is a digital rights technology company, enabling the fair and transparent use of copyright online. With Pex’s advanced identification technology customers can simplify licensing, brand safety, and compliance at the speed and scale of the Internet.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2014, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>81</td>
</tr>
<tr>
<td>Latest round</td>
<td>February 2021, Series A ($57M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$64M from Susa Ventures, NextGen Capital, Tencent, Illuminate Ventures</td>
</tr>
</tbody>
</table>

### Reality Defender

Reality Defender is dedicated to deepfake detection, offering a platform that scrutinizes user-generated media for falsified items, including voiceprints, doctored documents, and provides alerts for deepfake threats. Their service is primarily aimed at enterprises, platforms, and government bodies, utilizing advanced synthetic media detection algorithms. The company’s technology enables the real-time identification of manipulated content, aiding enterprise and government clients in combating synthetic alterations in images, videos, audio calls, and documents.

<table>
<thead>
<tr>
<th>Founded</th>
<th>2018, U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current headcount</td>
<td>35</td>
</tr>
<tr>
<td>Latest round</td>
<td>October 2023, Series A ($15M)</td>
</tr>
<tr>
<td>Raised</td>
<td>$15.97M from Y Combinator, Pioneer Fund, DCVC, ex/ante</td>
</tr>
</tbody>
</table>
The Current State of Content Moderation

Training general content moderation models for bulk content moderation can be costly because it requires significant developer resources and training data or a costly cloud subscription. Further work is also required to ensure machine learning–based content moderation outcomes align with platform policies. Moderation actions that require more nuance than a general model also add significant complexity and costs. Advanced models require specific training datasets, implementation expertise, and constant iteration due to the ever-changing nature of online harms. At present, content moderation BPO services often outperform machine learning systems on a cost–adjusted basis. Therefore, companies heavily rely on BPOs for content moderation despite an apparent appetite for software solutions.

The diagram below is a general representation of how T&S BPO Services and Software content moderation solutions both operate within a T&S organization:
However, with more innovation in AI and big data analytics, as well as the emergence of standard content moderation guidelines, experts anticipate that spending will shift from BPOs to T&S software options. Founders with T&S expertise are building towards an inflection point where automated solutions are becoming more precise than manual reviewers and costs for automated solutions drop significantly.

In the near term, there is an opportunity for new vendors to improve the organizational effectiveness of BPOs as well as support technology platforms’ in–house T&S activities. As T&S vendor capabilities continue to improve, experts predict that SaaS moderation platforms will begin to displace BPOs as the “T&S command center,” and AI–driven specialized tools will further support core capabilities, like risk intelligence products that identify and evaluate risks and trending events to shape T&S policies, purpose–built AI classifiers that detect problematic content at greater precision and recall than a human reviewer, and more.

The Future of T&S
While many T&S solutions aim to accurately apply policy–directed rules across various types of content, generative AI models have demonstrated an increasing skill in parsing the meaning of complex rules and applying them to nuanced content. T&S leaders are bullish on this emergent capacity as a key supplement to traditional content moderation, and even as a replacement in the long term. Effective integration of new AI tools may also reduce the need for human review of the most harmful content, mitigating psychological damage on many thousands of content moderators.
In-House Content Moderation Opportunities

Crucial for many in-house T&S teams, many generative AI tools may be combined with existing systems and stacked in ways that increase efficacy. For example:

- LLMs that leverage Massively Multilingual Speech (MMS) models to extend moderation across languages.
- LLMs that audit and refine existing machine learning classifiers or enrich regular expression keyword lists.
- LLMs embedded in segment/network analysis to surface content trends and detect new risk areas.
- LLMs that analyze and A/B test new content policy proposals, or even suggest new ones after reviewing edge cases.

Even at highly-resourced Tech Giants, months often pass before operations teams can fully implement new content policies, given the time needed for human labeling and classifier testing. But LLMs appear primed to collapse that timetable by minimizing the need for new classifiers and supercharging testing at scale in a sandbox environment. Finally, some believe generative AI tools will allow more opportunities for community governance as an alternative to one-size-fits-all content policies, reducing dependence on centralized policy teams.

Other Opportunities

Beyond content moderation, T&S teams appear likely to integrate generative AI capabilities into backend platform integrity systems. Particularly in the case of social media, many systems that improve ecosystem quality are not user-facing, ranging from feed ranking algorithms to age gating and behavioral analysis, among many others.
Appendix D
Drivers of Market Growth

As platforms become more integrated into daily life, the marketplace of solutions that address affiliated T&S issues will continue to evolve. Since failing to address T&S can be catastrophic for platforms operating in complex digital environments, experts anticipate that market growth will continue to accelerate, influenced by both business pressures and macroeconomic factors.

Business Pressures
Platforms undertake activities related to T&S for the following business and social responsibility reasons:

1. **Maintaining brand safety**
Platforms must be able to ensure to advertisers that their ads are not associated with inappropriate content that could damage a brand’s reputation. As advertising on social media has become mainstream and a main source of revenue, advertisers have banded together to make decisions about where to advertise and where not to and avoid the risk of their brands being shown next to harmful content. Disney, for example, pulled their advertising spend from X when they were concerned about their ads being shown next to antisemitic content in 2023. Associations such as the [Global Alliance for Responsible Media (GARM)] by the World Federation of Advertisers aim to see less harmful content online with more oversight and tools and have been one example of this type of industry leadership.

2. **Media scrutiny and increasing consumer expectations**
Effectively managing content- and conduct-related risks improves public perception and mitigates negative press coverage. Media scrutiny has also led to increased consumer expectations for an improved user experience with more robust T&S safeguards and quality content – if safety measures do not meet consumer expectations, consumers have a rising number of platform alternatives to take their business and attention to. The proliferation of Twitter alternatives immediately after the

---

16 Formspring, an early success of the anonymous online Q&A genre with 1M users within 45 days and $14M raised, hosted hateful content that led to a series of teen suicides and subsequent boycotts. While the site eventually added features designed to discourage bullying, the [site shut down] within 4 years.
17 A [workshop] with T&S professionals and academics found that most platforms heavily weight content quality metrics in their algorithms to increase long-term user retention.
18 71% of respondents to the [Edelman Trust Barometer] believe brand “CEOs are obligated to pull advertising money from platforms that spread misinformation.”
19 83% of consumers [stated] that tech companies are not doing enough to promote trustworthiness in their communities.
20 As a result of prioritizing posts containing “knowledge and advice”, LinkedIn has seen a [80% reduction of user complaints] about irrelevant content on their feeds.
takeover of Twitter by Elon Musk is illustrative of this. In 2023, disaffected users began to flock to alternatives such as Mastodon, Threads, and BlueSky.

The pressure for advertisers to assess brand safety questions and heightened sensitivity to reputation risk is likely to increase, rather than decrease. More than in years past, there is broader consumer awareness about risks on platforms due to major offline events that have been impacted by social media: elections, the COVID-19 pandemic, and large-scale violence, to name a few. The widening landscape of consumer platforms means that there are also immediate real world implications with marketplace services like Airbnb, Uber, and Rover.

3. Improved industry standards
There is also a more defined culture around safety among T&S employees at platforms. There has been a social transformation in the people who work on trust and safety over the last several years, elevating people from anonymous labelers and reviewers to leaders who are helping set new standards for quality, rigor, ethics, and human rights commitments. The professionalization of the industry through organizations such as the Trust and Safety Professional Association, the Fair Play Alliance (gaming), and Coalition for Trusted Reviews (e-commerce) are developing norms around best practices. This has led to changes in the ecosystem: it has required companies to pay attention to the investments they are all making; raised the bar for new entrants and made ad-hoc, at times invisible work, more visible and necessary. Industry organizations like the Integrity Institute are also developing policy memos to advise on upcoming legislation on trust and safety issues.

4. Emerging regulation & compliance
While T&S has largely operated in a voluntary self-regulation zone in the past, the E.U. and U.S. government, among others, have been increasingly defining legislation on T&S issues over the past decade. National and local regulations have, for the greater part of T&S history, driven the direction of this work. Companies have quickly become adept at detecting content that violated copyright laws, CSAM, or promoted terrorist content, developing new internal tools and teams focused on surfacing and taking appropriate actions on this type of content. The more these investments were able to help companies adhere to applicable laws, the better – a failure to do so could result in a company’s services being blocked or restricted.

The latest Freedom on the Net report highlights that at least 48 countries have pursued new rules for technology companies in 2022, with half relating to how platforms treat content.

The growth of platforms with user generated content and their rising influence have led to regulators taking action, particularly in the past 12-18 months. Companies operating in certain countries will be required to manage specific aspects of trust & safety, particularly relating to handling “content- and conduct-related risks to users and others, and mitigating online or other forms of technology-facilitated abuse.”

Platforms will need to leverage new products and streamline operations in order to achieve compliance with active and proposed regulations, which covers complex issues ranging from
managing content related to terrorism, child sexual abuse material and more, age verification, producing transparency reports, and new user controls, among others, across a wide variety of geographies. While large platforms may be able to build these management and reporting tools in-house, third party vendors will be integral to ensuring all eligible companies are able to comply.

Notable examples include:

Laws that have been passed:

**United Kingdom**
- **U.K. Online Safety Bill**: The U.K.’s Online Safety Bill, passed in October 2023, dictates how platforms must handle content like terrorism and child sexual abuse material, their obligations around child safety, age verification, producing transparency reports and more. Failure to comply will result in fines.

**European Union**
- **E.U. Digital Services Act**: The Digital Services Act (DSA), which will be applicable to all services available in the E.U. in February 2024, obliges services to adopt effective processes to address illegal content, such as hate speech and strengthen transparency of platforms’ decision making. The DSA applies to providers of digital “intermediary services” and obligations are tiered, with lower-level obligations for online platforms, like marketplaces and app stores, and advanced obligations for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) that reach over 45M monthly active EU citizens per month.
- **E.U. AI Act**: In December 2023, the E.U. passed the E.U. AI Act which prohibits AI systems that reach an unacceptable risk to people's safety, including those that employ subliminal or purposefully manipulative techniques.

**Australia**
- **Australia Online Safety Act**: Passed in June 2021, the Australian eSafety Commissioner requires platforms to produce compliance reports with Basic Online Safety Expectations as defined by the Australian Government. Following each report, the eSafety office has published summaries and commentaries of this information (unlike the DSA, whose transparency reports are published by the platforms themselves).

**Singapore**
- **Singapore Online Safety Act**: Passed in February 2023, this added a new section to the Broadcasting Act that enables the online safety regulator, the Infocomm Media Development Authority, to ensure businesses comply with a code of practice around specific harms, including terrorist content.

**Proposed laws:**

**United States**
Platforms operating in the U.S. are provided broad protections via Section 230 of the Communications Decency Act, which removes their liability for hosting, removing or limited access to user generated content. However, regulators have proposed numerous bills and hosted hearings to
investigate the role of platforms in disseminating misinformation and extremist ideologies, as well as understand how regulators can help contain online harms:

- In February, 2021 Democratic senators proposed the **Safeguarding Against Fraud, Exploitation, Threats, Extremism and Consumer Harms (SAFE TECH Act)**, which would make online service providers liable for specific activities related to cyber-stalking, targeted harassment and discrimination.
- In 2023, Sens. Blumenthal and Blackburn proposed the **Kids Online Safety Act** to ensure that platforms put the interests of children first by giving minors and parents options to protect their information, creating accountability via audits and other forms of compliance, and facilitating algorithmic transparency, among others.
- The White House **secured voluntary commitments from leading artificial intelligence companies in July 2023**, which requires that companies commit to developing robust technical mechanisms to ensure that users know when content is AI-generated, such as watermarking. The subsequent **Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence** notes the potential for irresponsible use of AI to exacerbate harms like disinformation.

**India**

**Digital India Act:** The **Digital India Act**, put forward in early 2023, is designed to replace the Information Technology Act of 2000 with a legal framework that keeps pace with the country’s burgeoning digital economy, anticipated to reach $1 trillion by 2026. India boasts the greatest number of Internet users globally, with 850 million and growing, and the updated regulations include key components related to online trust and safety. This wide-ranging legislation encompasses safeguarding users against cyber threats, such as revenge porn, defamation, and cyberbullying, as well as protecting minors and their data from addictive technologies and fake news. The Act will require a review of the country’s ‘safe harbor’ principle, which removes social media companies’ liability for user generated content, and introduces new user data protections.

Additionally, platforms must fulfill their terms of service, which might include age verification, and uphold their community standards of what is and is not acceptable content/conduct.

**Macroeconomic Tailwinds**

The Report also identified the following **macroeconomic tailwinds** for the T&S solutions market:

1. **Geographic Expansion**

   Global expansion of platforms is a key growth factor for demand for T&S solutions because revenues from digital advertising (social media and other channels), E-Commerce, platforms (ride-hailing, dating, delivery), and mobile applications (in-app purchases and paid apps) underpin T&S budgets. Therefore, growth in the population of active Internet users and their spending power cannot be ignored.

   According to survey data from the International Telecommunications Union and World Bank, **59.6%** of the global population actively uses the Internet. Approximately **73%** of the global population is
between age 10 and 65, the target demographic for internet adoption. This suggests that there is an untapped market of nearly 1.07 billion additional internet users at present. Most non-internet adopters live in growth markets, mainly South Asia, Central/East Asia, Sub-Saharan Africa, Latin America, and the Middle East. Therefore, the development of these traditionally underserved regions represents a significant opportunity for digital services and advertising revenue growth. Capturing these growth markets will increase T&S spending as new users join internet platforms.

2. Digitized Society
Digitization refers to the integration of digital platforms into everyday life. Market penetration statistics for food delivery, use of social media for work purposes, e-commerce adoption, and other markets indicate that the diversification of online platform applications and their increased importance in everyday activities will fuel the need for T&S solutions to manage these new mediums.

In 2023, 4.8 billion social media users were identified by Datareportal, and the average social media user engages with 6.6 platforms each month.

Unsurprisingly, global social media adoption closely follows global expansion trends, with the greatest opportunities for growth in developing markets.

3. Outflux of Trust & Safety Talent
Founders are taking the nuanced expertise they honed addressing T&S issues at the major platforms serving billions of users and building new T&S software startups. A combination of T&S professionals noticing gaps in the T&S market coupled with technology industry layoffs have led to a major outflux of talent from companies with acute T&S issues and sizable budgets to address them at scale. For example, Meta’s contract with Accenture for outsourced content moderation activities is valued at $500M annually.

At present, an abundance of highly skilled T&S experts are founding and supporting startups that were created specifically to manage T&S risks. Teams’ deep and practical understanding of the problem space should help them define the category and attain product-market fit.

According to CNBC, the founding team at T&S tooling startup TrustLab, which raised ~$23M, came from companies like Google, Reddit, and TikTok parent Bytedance, while the founders of Intrinsic, raised $3.6M, worked on T&S issues at Apple and Discord.

SaaS moderation platform Cove, which raised ~$6M, was founded by software engineers from Meta and Twitter, while Cinder, which raised $14M, assembled a founding team from Meta’s threat intelligence group. According to Cove CEO Michael Dworsky, “I think the cost cutting has definitely obviously affected the labor markets and the hiring market. There are a bunch of brilliant people out there that we can now hire.”

Key Considerations and Challenges
However, market concentration and software efficiency were identified as headwinds for the T&S market.

1. Market Concentration

While the SAM analysis for this report tracks 261 total companies, roughly two-thirds (66%) of estimated spending comes from the largest companies in the Tech Giant segment. T&S tooling M&A activity has also picked up over the past 5 years, and experts have seen bigger platforms or other vendors acquire as many as ⅓ of all T&S software startups in the past decade. This time period was defined by an increasing market need for T&S talent, both on the vendor side and by platforms, cash on hand for companies to acquire startups, and a point of technological development before the risks of genAI were clearly articulated by the public. While this activity may decrease the number of vendors available on the market for smaller platforms and consolidate tooling spending with large vendors, T&S software startups are primed for growth today with more acknowledgment from platform customers.

2. Software Efficiency

History has shown that technology is deflationary – successful development of T&S technology will allow T&S teams to keep more users safe at a lower cost. Therefore, future gains in market value for software will likely come at the expense of some services spending. While a shift from expensive

---

Companies like Sentropy, which raised $13M from Initialized Capital and more, were acquired by Discord in July 2021, just a year after they launched. In October 2022, Spotify acquired SaaS moderation company Kinzen after they raised ~$3M the year prior. ActiveFence, the SaaS moderation platform with $108M+ in funding, acquired Spectrum Labs in September 2023 after they raised $32M from Intel and Greycroft.
outsourced human moderation to lower-cost software tools represents an advancement in the T&S industry, it is likely to put deflationary pressure on overall levels of spending.

Additionally, **BPO employee unionization** is another key consideration that platforms are managing, as content moderators worldwide are raising alarms about their poor working conditions and exposure to traumatizing content. Accelerating in 2023, these workers have been **unionizing** in regions ranging from Africa to the E.U. In Kenya, Meta was required to pay $52M and $85M to more than 10K subcontracted moderators in two settlements to compensate them for psychological harm sustained on the job. In Ireland, where BPO employees are protected by trade unions, mass layoffs conducted by Accenture were met with **inquiries from their professional representatives**. As partnerships between BPOs and platforms become less predictable, enterprises may be more willing to try new solutions that can fulfill their needs.

**Drivers of Trust & Safety Budget (and Vendor Needs)**

A few predictable forces can influence T&S budget. The timing between the onset of these drivers and the downstream effect on budget and vendor decisions will vary depending on the flexibility of the company’s budget management processes as well as the severity and timeliness of the driver.

**1. Ad revenue & user satisfaction:** If the company relies heavily on ad revenue, then brand safety and a robust, well-funded T&S program are necessary to maintain advertiser trust. Furthermore, if user satisfaction is highly important to the company’s success, this will also drive up T&S budget, as customers (**60% of U.S. adults**) consider a brand’s reputation, values, or ESG position as the “most important” factors when making a purchase.

**2. Crises & Public Relations:** Regardless of whether it becomes public, crises create an urgent need to expand T&S capabilities. Crises often reveal gaps in proactive detection and enforcement, which take time to close through product solutions and operational changes. Vendor teams can be a useful stopgap measure for companies in these moments. A few factors will raise the stakes of a T&S crisis (and the potential need for time-sensitive and expert solutions): (1) if the risk and/or impact are novel to the company, (2) if there is potential or actual negative PR.

**3. Compliance:** Complying with (often) complex regulatory requirements, whose scope and timing are out of a company’s control, can represent a daunting undertaking and detract from achieving the pre-existing priorities of the company and its T&S org. This is especially true for smaller companies whose staff isn’t large enough to absorb and respond to exogenous factors and where there is less institutional experience with regulatory requirements. Regardless of company size, vendors may represent a fast and assured compliance solution; vendors can serve as a hub for institutional regulatory knowledge, having solved many (similar) compliance requirements across multiple clients. The need for vendor support for regulatory compliance would be further amplified if the regulations involve aggressive deadlines for adherence.

**4. Product launches; market expansion:** Larger audiences and/or new products can meaningfully expand a company’s exposure to T&S risks and internal teams and capabilities may not be able to expand on the same timeline. For example, if a company can confidently detect T&S risks for the U.S. market via English language classifiers built by their California-based engineers, expansion into
Mexico and Brazil would require those technical systems to be internationalized with a diversity of cultural and language examples and data. Enforcement operations would also need to be augmented with Spanish and Portuguese speakers. Such product and market expansions would benefit from vendor support who offer globally diverse data, as well as topical and language expertise.

5. Investment in T&S (flywheel effect): Investment in T&S (across headcount, engineering priorities, vendors, etc.) can beget more investment in T&S. When a company launches a comprehensive effort to accomplish a particular T&S goal, the effort may lift the hood on a broader system of T&S risks. This often reveals a more complex picture and/or new, previously unknown T&S risks, which then will require even more attention and can potentially benefit from vendor support. For example, while investigating the users behind posts that propagate the same policy-violating image, the company may discover that the same users are involved in a previously undiagnosed form of abuse whereby they organize efforts to harass local politicians on the platform.

6. Risk of Harm: Demand for T&S support will increase as the risk (or actual manifestation) of real-world harm increases. For example, a platform that only publishes highly vetted content (e.g., Coursera) is less likely to encounter the risk of real-world harm than a scaled platform (e.g., TikTok) or platforms that facilitate in-person interactions (e.g., rideshare and dating apps).

Considerations for Resourcing Decisions: In-House or Vendor?

There are several T&S functions (as outlined in the T&S solutions stack) that are especially well-suited for such vendor support based on the overlap between T&S organization needs and the value that T&S vendors can deliver. However, the below considerations should be applied to validate and possibly expand upon this list to ultimately determine when it's most appropriate to engage vendors:

Risk Readiness: Is the company ready to assume the risks and obligations associated with handing the T&S issue(s)? For example, safely engaging with certain types of content (e.g., child safety), requires the company to adhere to extremely strict regulations as well as legal and law enforcement obligations. If not, vendors have the experience and knowledge to appropriately handle such cases.

Budget & Hiring: Vendor support may be more effective for solving a T&S issue with regard to cost and/or time to hire. For example, licensing an existing vendor tool is likely to be more cost effective than building the solution internally. In addition, vendors can provide flexible or short-term access to subject matter experts, thus obviating the need to hire for that expertise.

Timeliness: How time-sensitive is the risk? Speed is often important in addressing T&S issues, so unless the company already has the necessary knowledge and well-suited systems/tools, vendor support is likely the more timely vs. building something from scratch.

Suitability of Institutional Knowledge & Tools: How effective are in-house tools? How relevant is in-house knowledge? (e.g., familiarity with the subject matter of the content, ability to understand bad actor tactics, knowledge of the regulatory requirements). If the company cannot adapt existing systems or deploy existing knowledge, vendor solutions may be beneficial.
**Data & Privacy:** In order for a particular vendor tool to be effective, would it require that the company share proprietary data, IP, or user data? Is there an avenue to make use of vendor tools without compromising such sensitive data?

**Bias:** Within the company, is there sufficient diversity of opinion and expertise to comprehensively address a topic in an unbiased fashion? Vendors often bring in a diversity of global data and culturally sensitive perspectives.

**Wellness:** Is the company equipped to understand and support the wellness of in-house staff who are exposed to problematic content? Some companies deem it appropriate to task vendors with content moderation given their ability to provide wellness support.

**Alignment with Company Mission & Priorities:** Does building in-house T&S solutions represent a diversion from the company’s priorities? When faced with limited resources (funding, time, talent etc.), a company may find that vendor solutions are preferable to pulling internal resources away from other priorities.

**Company Maturity:** Does the company’s maturity and availability of resources lend itself to building in-house? It’s likely that early-stage companies with fewer staff and resources will frequently use T&S vendor support to create a comprehensive T&S system, while enterprise-level companies with dedicated T&S resources will seek T&S vendor support for more specialized use cases.
## Appendix E

### T&S Team Overview

Because T&S teams at platforms are relatively nascent and evolving quickly, there is variability across companies around who is ultimately responsible for managing T&S risks and the associated budget for these purchasing decisions.

While the details of an organization will vary, T&S teams generally fulfill three key functions, which differ in how they allocate resources:

<table>
<thead>
<tr>
<th></th>
<th><strong>Policy Development</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>These policy teams are deep subject matter experts on problematic content and behavior that could be (or already is) present on the company’s platform. They are driven by principles and design rules that can be applied at scale to maximize the prevention of harm. They review a handful of representative examples of risky content and behavior for the purpose of informing each new policy.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><strong>Policy Enforcement</strong></th>
</tr>
</thead>
</table>
| 2 | Intelligence and Investigations teams proactively monitor trends and research risky content and behavior, while operational teams detect and review this content and behavior in order to take action according to platform standards (e.g., remove the content, disable user access, etc.)
Policy enforcement teams will also facilitate escalations if addressing the content/behavior requires participation from other teams or external entities: for example, coordinating with law enforcement on illegal content, or sharing information with security teams if content presents a physical risk to employees. |

<table>
<thead>
<tr>
<th></th>
<th><strong>Technical Enablement</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Technical teams support policy and operations teams by building systems (e.g. classifiers to identify large bodies of risky content) and tools (e.g., case management) that help to detect and enable action on problematic content/behavior. They may also build tools that discourage or prevent bad behavior, for example: systems that flag potential policy violations before they publish content or comments, or systems that require users to passively meet a series of benchmarks (e.g. 98% of posts without policy violations) before they can unlock the more sensitive features that carry larger risks (e.g. live streaming access). T&amp;S-specific technical enablement teams are likely to be robust in large enterprise companies, but early stage companies are unlikely to hire product managers or engineers solely dedicated to T&amp;S.</td>
</tr>
</tbody>
</table>

---

22 Consistent exposure to troubling content is known to have a negative effect on these teams’ mental health and wellness. It’s important to note that discussion of and exposure to problematic content affects everyone differently; supporting mental health and wellness should be a priority for all T&S teams.
As technology formats evolve, companies with a digital footprint will seek those with T&S policy, operations, and technical experience to manage emerging risks. In the age of AI, those with T&S experience are already being tasked to design and lead safety efforts within AI companies themselves.

**Who Makes (Vendor) Budget Decisions within Trust & Safety Orgs?**

The need for vendors differs across these three functions, as does the power to make budget decisions to enable such vendor support. The process of defining a budget will vary across companies, but by and large, policy enforcement teams will have the largest demand and budget discretion for vendor support, primarily due to the fact that they are on the front lines of addressing ever-changing T&S risks, require extensive human and technological resources to moderate these risks, and face internal and external pressure to do so.

Demand for vendor-based technical solutions (automated detection, classifiers, case management systems, etc.) will vary based on the maturity of a company’s T&S technical enablement teams, but the policy enforcement teams are likely to be the advocates for any vendor coverage to fill technical gaps, as the enforcement team directly feels the effects of any deficiencies in systems or tools.

In order to keep pace with emerging trends and understand new topics with the depth required to design new policies, policy development teams will often seek external T&S services and tools. Vendors help policy development teams understand and stay ahead of risks by providing landscape analyses (e.g., how geopolitical factors impact content risks), delivering deep-dives on specific risks/trends (e.g., state actor behavior on the platform), and providing access to topic-specific expertise that can inform policy development (e.g., a firearm expert helps internal teams understand which elements of gun-related tutorials are dangerous). Recently, policy development teams have seen increasing regulatory and civil society pressure to engage external experts in the policy development process.
Appendix F

Corporate Benefits of T&S Vendor Partnerships

How T&S Vendors Enhance Business Operations

Vendors can benefit business operations across efficiency, flexibility, and efficacy. Below is a non-exhaustive list of examples and possible key performance indicators (KPIs).

Efficiency:

- **Prioritization**: Internal teams can optimize their time when vendor solutions prioritize issues by severity and group similar issues for bulk review. Similarly, teams can better prioritize policy development efforts when vendor landscape analyses present a comprehensive view of a risk area.

- **Complete and organized information**: Vendor tools organize data and issues so that internal teams have all necessary information in front of them in order to evaluate and make decisions (vs. needing to piece together disparate data).

- **Off-the-shelf solutions**: Licensing a proven vendor solution saves time vs. building a solution internally from the ground up.

- **Subject-matter expertise**: Vendors’ industry and topical expertise allows them to recognize issues and develop solutions more quickly than internal teams who may be seeing an issue for the first time.

*Example internal KPIs:* time to review and/or take action on issues; cost per review; time saved by internal teams (for gathering data, by outsourcing tasks to vendors).

Flexibility

- **Flexible staff**: Vendors provide access to a diverse pool of experts, as opposed to the often lengthy process of hiring full-time internal team members with less fungible expertise.

- **Flexible timing**: Vendor solutions offer comprehensive and round-the-clock availability, whereas internal capabilities may be limited by time zone or on-call availability. Vendor staffing can respond to timely system issues (e.g., outage of an important enforcement dashboard) instead of reassigning internal staff or stretching their bandwidth.

- **Short-term support**: For discrete and/or time-bound issues, hiring temporary subject-matter experts can be preferable to developing internal expertise.

*Example internal KPIs:* % of time with T&S coverage (e.g., 24/7); overtime required of internal staff; time to cover niche issues.

Efficacy
• More informed delivery and decisions: In cases where vendors possess more subject matter expertise than internal teams, vendors are able to more effectively evaluate content, identify risks, and deliver work products. A vendor may have solved a particular problem several times over with other clients, while a company tries to wrestle with it for the first time. Separately, a vendor’s solutions may identify opportunities to improve the efficacy of internal work by providing data that measures impact. For example, a vendor’s case management tool could identify that a certain process is occurring slower than the norm.

• Access to cross-industry insights: Vendor solutions incorporate lessons from a broad landscape of T&S risks observed across their numerous platform clients. Incorporating such cross-industry solutions is likely to strengthen an internal–only approach to T&S. For example, a vendor’s proprietary solutions may use publicly available data across multiple platforms to identify risk. Or, a vendor’s deep understanding of problematic behavior on Platform A can give them a leg up in designing and implementing solutions to detect similar problematic behavior on Platform B.

• Effectively protecting internal staff: In some cases, companies may determine that certain tough issues are most effectively handled by a vendor due to the vendor’s experience with providing wellness support to their staff. Vendor tools can also protect internal team members from troubling details (e.g., via tools that omit or obscure irrelevant info) which allows staff to focus on the most relevant information.

• Counteracting bias: Incorporating vendors’ external expertise diversifies inputs into company policies and systems (e.g., classifiers) and can counteract bias.

• Expert validation: Utilizing vendors’ industry and topical expertise can provide validation for and/or increase confidence in internal T&S decision-making (e.g., industry validation that a threat actor poses the risk of real-world harm).

• Proactive and scaled impact: Vendors can help companies employ more proactive methods for addressing T&S risk. Vendor tools and insights can help companies identify emerging risks early enough to design a response (e.g., new policies, bulk actions) that can prevent the risk from propagating across the platform. Vendor tools and expertise can also help companies identify T&S issues at the source (e.g., a specific group of users responsible for the majority of the bad behavior) so that the company can take a systematic, scaled approach rather than addressing issues on a case-by-case basis.

• Precision and recall: Vendor solutions can improve a company’s ability to identify a greater portion of T&S policy violations (recall) as well as improve the company’s ability to differentiate between true T&S policy violations vs. allowable content/behavior (precision). For example, a vendor’s technology can expand upon a known policy violation to identify new derivations (e.g., content that uses masked language to avoid detection). Or a vendor could apply its cultural expertise to identify that a problematic English term does not have the same meaning in Spanish, such that content containing the Spanish term would not violate T&S policies.
Example internal KPIs: novel issues identified; actions taken per risk (e.g., content removed per risk identified); precision and recall; actions taken via proactive vs. reactive enforcement.

Revenue and User Growth Trends on Safe Platforms
It can be difficult to measure the direct, positive effects of a robust T&S program due to the nature of T&S: if the program is working well, there will be no observable issues because risky content and behavior are removed before anyone (users, advertisers) is exposed to it. That being said, there are a few predictable patterns whereby T&S programs foster positive outcomes.

Users
User Engagement, Retention, and Growth
When users are protected from harmful content, they are more likely to return to the platform, thus driving up daily active users and reducing churn. Said another way, as a user’s exposure to problematic content increases, they’re more likely to leave the platform. This is particularly important for platforms with younger users, as perceptions of safety heavily influence parents’ willingness to allow their children access to an online service. For example, in October 2023, after observing problematic Hamas videos on TikTok, Israeli and Jewish schools recommended that parents disable their kids’ access to prevent further exposure to harmful content.

When users feel secure, they are also more likely to engage in meaningful interactions (sharing content, commenting on posts, etc.) which contributes to a healthy community and a positive experience for other users. All of these user experience benefits (feeling safe, engaged community) will also serve to support user growth. Matthew Soeth, former head of community at Spectrum Labs, has noted that their clients have seen an increase of customer LTV by 30% by reducing harmful content and promoting healthy behaviors.

User Contributions: UGC, UGS, etc.
Creators will invest in content production for safe platforms, as these platforms are more likely to support steady, uninterrupted creator income due to fewer brand safety issues. Investment in T&S programs can also incentivize safe user contributions to the platform. For example, if the T&S program makes it easy for a marketplace seller to understand how their draft listing would violate T&S policies, the seller will be able to learn, adapt, and post this and other listings successfully, improving the overall user experience and reducing future violations.

Revenue & Funding
User-Driven Revenue (e.g., subscriptions)
In the same way that a safe platform drives user engagement, it will also fuel user-based revenue streams; higher user growth, user satisfaction, and user engagement enabled by a safe platform will drive subscription revenue and on-platform transactions. However, users with poor experiences will spend less money. According to ADL’s Hate and Harassment in Online Games 2023 survey, 20% of

23 Jeff Horwitz’s book Broken Code about Facebook reveals that “integrity work has boosted usage among new users by 0.9%” and overall usage by “a fraction of a percent”. 
adults and teens are spending less money within online games due to the hate and harassment they face on these platforms.

Advertising
Brands tend to spend advertising dollars on platforms that are aligned with the brand’s values, provide positive user experiences, and minimize risks associated with unsafe content; robust T&S programs help create these conditions for healthy ad revenue. When T&S programs pair ads with unsafe content, advertisers will often suspend ad spend to protect their brand. For example, the recent discovery of unsafe child safety content on Instagram led prominent dating apps (Bumble, Match) to stop advertising on the platform.

Funding
While not relevant for all platforms, funding opportunities may be influenced by the strength of a company’s T&S program, as some investors believe that this is an indicator of longevity. Michael Dworsky, CEO of Cove, has been approached by multiple companies pre-launch due to investor pressure for a T&S solution early in the platform’s life cycle.

Potential Cost Savings and Risk Reduction
Investing in T&S will allow the company to generate cost savings associated with efficiency gains, resilient employees, preventing regulatory costs, interruptions to focus, and brand safety issues. A strong T&S program will also help the company mitigate the risk of harm as well as address systemic platform risks.

Cost Savings
Efficiency, efficacy
A strong T&S program will enable both internal and vendor-based solutions that mitigate T&S risks effectively and at scale. For example, T&S classifiers that automatically remove problematic content will reduce the need for time- and labor-intensive human reviews. Vendor support for a T&S program can also create cost savings via the efficiency, flexibility, and efficacy gains. For example, ActiveFence’s ActiveOS tool has enabled teams to experience a 38% improvement in operational efficiency due to better decision-making, analytics, and prioritization tools. As a result, platforms can use the cost savings towards improving their current and future moderation capabilities.

Company wellness
Beyond the need to support wellness as a corporate responsibility, a program with effective wellness support for T&S employees will foster better outcomes including higher employee engagement, job satisfaction, productivity, and teamwork. Additionally, adopting T&S software to automatically detect content will decrease costs associated with moderator wellbeing. On the other hand, when T&S wellness support is insufficient, the company will be faced with the costs associated with lower productivity, absenteeism, and turnover.

Regulatory & legal costs
A T&S program that effectively detects and remedies issues is less likely to face the growing body of regulatory investigations and fines. As an example, the E.U.’s Digital Services Act (DSA) can impose fines of up to 6% of global annual revenue if a platform doesn’t sufficiently crack down on harmful content; further, a platform can be temporarily suspended in the E.U. if it continually refuses to comply. As legislation continues to develop around the world, the impact of T&S–related regulatory fines (in relation to the bottom line, and especially to smaller platforms) remains to be seen.

Finally, a T&S program that prevents issues at the source is less likely to need to pursue costly downstream legal remedies. For example, Google recently sued an alleged scammer for creating hundreds of fake companies and thousands of fake reviews.

**Indirect costs**
When a T&S issue occurs, any external scrutiny will dramatically increase the effort required to address the problem, which can disrupt the company’s focus. Regulatory, PR, and brand safety issues often require an all–hands–on–deck approach, which can pull resources and focus from other important workstreams (T&S or otherwise). For example, if Congress launches an investigation after a platform failed to protect users, in addition to the T&S teams who will directly remedy the issue, the platform will need to dedicate PR, legal, and CEO attention to respond effectively to Congressional hearings, written briefings, etc. Robust T&S programs will be better equipped to prevent and detect issues early and avoid the disruptions that result from external attention on T&S misses.

**Risk Mitigation**
**Reduce risk of real–world harm**
Robust T&S programs help platforms prevent real–world harm. Comprehensive detection, scaled enforcement capabilities, and knowledge that helps the organization stay ahead of evolving risks will all contribute to the company’s ability to remove content or behavior before it can cause real–world harm to user(s) (e.g., bomb–making tutorials, harassment messages, users dealing drugs).

By preventing problematic behavior and removing risky content, a strong T&S program will also make it difficult for bad actors to achieve their goals on the platform, and in such cases, bad actors will likely find that advancing their agenda on the platform is no longer worth their time. Making the platform unappealing for bad actors can reduce the aggregate risk for real–world harm by stopping it at its source.

**Reduce systemic risk**
Investment in T&S will build institutional knowledge that will allow the company to better anticipate and recognize T&S risks and address them at the system–level. For example, a robust T&S program will be able to identify common bad actor tactics and make connections between related risk areas, which can inform broad, platform–wide solutions. In addition, as described in Investment in T&S (flywheel effect), when a company tackles a specific T&S issue, the exercise often exposes a broader system of T&S risks, thus allowing the company to mitigate systematic risks rather than discrete issues.
Reduce risk of brand safety issues

Strong T&S programs will be able to detect and remedy issues before they impact users or advertisers, thus preventing brand safety issues. Even if a brand safety issue does occur and ads are paired with problematic content, a mature T&S program will help the company to expeditiously and thoroughly address the issue and minimize negative impact to users and brands.

In addition, companies with mature T&S programs can give advertisers more agency in safeguarding their brand: when a company understands how T&S risks manifest across different categories of content, they can empower advertisers to have more control over which content categories they advertise on.
Appendix G

Case Studies

Cybersecurity’s parallels with the T&S industry

T&S is an adjacent field to cybersecurity, which is focused on preventing exploits that breach technical systems and data.24 However, the evolution of T&S shares similarities with the trajectory of the cybersecurity market, which in 2022 was valued at $222B globally.

By best estimates, the cybersecurity industry started in the 1970s with the introduction of the first antivirus software and the self-replicating program, or computer worm. This was commercialized in 1987, with the launch of the first antivirus product and later the founding of McAfee, which was bought out in 2021 for $14B. In the 1990s and with the wider availability of the internet, network security threats increased dramatically, and firewalls and antivirus software were mass produced to protect customers. In the 2000s, the risks associated with data breaches and ransomware attacks resulted in business disruption, loss of customer loyalty, lawsuits, and regulatory fines, which led companies to prioritize implementing dedicated cybersecurity solutions.

In parallel, governmental bodies were conducting cybersecurity research and developing guidance for the public sector and private companies, accelerating cybersecurity market expansion. For example, in 1974, the National Institutes for Standards and Technology published the first guidance for corporate executives on computer security, followed in 1977 by the first guidance on authenticating users. Standards created in the 1970s and 80s, such as a password usage standard and Data Encryption Standard were widely adopted in the private sector. Additional frameworks and standards, which included voluntary and mandated compliance, such as SOC 2 and ISO 27001, were also developed to help organizations improve their cybersecurity and information protection efforts.

It took years for an organizational function to evolve that was solely dedicated to managing cybersecurity risks and mitigations, with the role of Chief Information Security Officer becoming more

24 See International State of Safety Tech 2023, p. 9 for a further breakdown of how T&S is differentiated from cybersecurity.
widely adopted in the early 2000s, more than 20 years after the introduction of the modern internet.

In 2004, the global cybersecurity market was worth $3.5B. In 2023, the global market is valued at $249.8B, representing a 7039.43% increase over ~20 years. A 2014 Wired article quoted Peter Singer, who saw the industry’s growth mirroring that of the internet: "It will continue to grow, most likely on an exponential curve because it’s going to follow the internet itself. If five billion new people are coming online, five billion new security problems are coming online.

In 2014, Finnish cybersecurity expert Mikko Hypponen predicted that cybersecurity would have profound implications for military and defense, calling it the greatest “technological shift since the second world war.” He added, "We are entering an era where the big shift over the next 50 or 60 years will be the development of cyber-ops, completely virtual arms. It’s the beginning of the next big shift for the militaries. Fewer than ten countries have nuclear arms. Every single country, in theory, can have cyber arms.”

Market drivers that led to the hypergrowth of the cybersecurity market, such as business pressures, new regulations, a focus on national security, and technological advances, are manifesting today in new forms to contain T&S risks for platforms with user generated or genAI capabilities. In the coming years, the development of frontier models and proliferation of AI systems will create, to paraphrase Singer, “five billion new T&S problems as five billion people are using AI technologies.” Artificial intelligence is expected to deliver capabilities that will weave technology more intimately and tightly into daily life across critical industries and geographies, and the growth of the T&S solutions market will help augment the benefits of new technologies while containing known and unforeseen risks.

**Minecraft**

*How T&S can positively impact business performance*

Minecraft is an undirected worldbuilding “sandbox” played by [174M](https://www.mojang.com) monthly users across 3k servers in multiple platforms. It was created by Swedish developer Mojang Synergies AB (Mojang), which Microsoft acquired to build its gaming portfolio for $2.5B in 2014[26]. With more than 300M unique copies sold as of late 2023, Minecraft is the best-selling game of all time.

Millions of children form a significant base of Minecraft users. A [report](https://www.nielsen.com/corp/pdfs/2020/nrm_20p_20mcdonalds_2020.pdf) estimated 20% of users were under the age of 15, while 43% were between 15–21 years old (the largest demographic). A robust ecosystem of third-party experiences “from community–made skins and textures to hand–crafted worlds and epic adventures” are available on Minecraft Marketplace. Gameplay is uniquely visible. YouTube reported more than 1 trillion views on gameplay on its platform in 2021, with [Twitch adding](https://www.twitch.com) substantially to that figure.

---

The product’s reach, youth of its players, integration of third-party experiences, and heightened visibility of gameplay all combine to make effective T&S crucial to Minecraft’s business success.

“We goal our teams on PR [public relations] issues,” said one Minecraft executive in 2022, “that’s what we consider to be the minimum bar for success.” The firm has not always achieved that goal. The general surge in screen time during COVID-19 not only spiked Minecraft use, but also spurred alarm over child predation. The FBI warned parents and guardians to combat this threat, and a community of 120k parents protested what they saw as insufficient protections on Minecraft. Mojang responded citing parental controls, chat filtering, and the requirement that parents authorize downloads by children under 13. Minecraft’s general success with parents and game raters has mitigated concerns, receiving a Pan-European Game Information rating for children ages 7+ and a Common Sense Media rating of four stars for children ages 8+.

Variations in game versions, variety of platforms, and numerous servers make it difficult to assess T&S deployment across the ecosystem. But Minecraft has developed and deployed key security features including in-game player interaction controls, screening of Minecraft Marketplace partners, player reporting tools that leverage community standards across Minecraft and Xbox, and human review (with appeals). T&S enforcement does not merely respond to user-generated reports, however, and data released from Microsoft’s Xbox showed a proactive enforcement rate of 95% for the first half of 2023. Those statistics would not cover gameplay on platforms that do not leverage Xbox’s safety infrastructure, but Minecraft has found success in branding its T&S focus, announcing a third interactive game-based lesson equipping players with player safety skills on Safer Internet Day 2024.

T&S will doubtless remain a major focus for Minecraft, given the popularity of player modifications and associated vulnerabilities. But Minecraft has become a gold standard for demonstrating responsibility across a range of T&S to win parental approval, opening access to major youth markets worldwide which will continue to sustain its best-selling status. As a result, experts estimate that Mojang’s T&S solutions will grow approximately $5.8M at a CAGR of 9% by 2028 if it sustains its best-selling game status and remains a community gathering place among young users.

**Airbnb**

*How T&S can positively impact business performance*

**Airbnb Inc.** (ABNB • NASDAQ) has successfully pioneered peer-to-peer accommodations in the sharing economy, overcoming competition with incumbents that stand to gain from safety incidents, which could invite regulation and deter both hosts and guests. Highly resourced hospitality alternatives—including hotels, motels, and inns—and politically connected homeowner associations raise the stakes of public safety failures and put pressure on Airbnb to adopt and publicize safety measures that invite trust from the hosts and guests across the platform’s global customer base.

---

27 Xbox deploys AI-enabled filtering technology called Community Sift from T&S SaaS moderation platform Two Hat, which Microsoft acquired in Oct 2021 for an undisclosed amount.
The platform must both preempt regulatory intervention and appear responsive to offline community concerns that range widely from neighborhood disruption to unauthorized surveillance, property damage, health hazards, and even human trafficking. But Airbnb’s T&S program must also mitigate threats more common across online service providers: secure payments and refunds, authentic customer reviews, scam prevention, and customer data storage.

“We believe that our revenue growth depends upon . . . our ability to build and strengthen trust and safety on our platform and among members of our community,” Airbnb wrote in 2023, further describing the combination of programs the company calls its “system of trust.” Initially, T&S prioritization at Airbnb appeared to be reactive—a challenge common across industry:

- **Formation** of the T&S department in 2011 followed a Bay Area vandalization
- **Party bans** followed violent fatalities
- **Satisfaction guarantees** followed publication of an national scam
- **Anti-discrimination commitments** followed an indicting study

But Airbnb seemed to shift their T&S narrative to be more publicly proactive in 2022, announcing the formation of a Trust & Safety Advisory Coalition (TSAC) composed of “22 respected organizations” that would “advise on Airbnb’s community policies that govern the platform, as well as products,” among other things. TSAC has since contributed to a safety feature for individual travelers, a pool safety program, and membership in a multi stakeholder anti-trafficking alliance called The Code. In 2023, the firm also profiled and praised the women leading its safety work.

A review of Airbnb’s T&S efforts generally shows safety announcements and increased publicity have coincided with the company’s initial public offering in December 2020 and growth of the firm overall, but based on interviews with former employees, the company has always had an ethos of proactively designing for trust. This culture of trust is reflected in their branding of key initiatives like AirCover, Solo Traveler Safety, Airbnb Community Commitment, and Get-What-You-Booked Guarantee, and in their November 2018 acquisition of T&S SaaS moderation platform Koko.

Airbnb’s evolution from novel peer-to-peer platform to major force in the hospitality industry was enabled by the unique combination of positive customer experiences and essential T&S offerings. Their strong business performance demonstrates the essential value of investment in T&S programs and policies, as well as the need to reinforce public confidence by branding and publicizing those efforts. In the next 5 years, experts estimate that Airbnb’s T&S solutions budget will grow approximately $47M at a CAGR of 8.5% if it grows in market dominance against hospitality incumbent competitors.

**X (formerly Twitter)**

*How a divestment of T&S impacts a platform's market positioning*

---

28 [IOK p. 7, 11](#)
In 2023, advertisers were displeased with X’s new T&S policies – a development that has impacted its revenue and positioning in the market. For X, a decline in advertiser trust has become an existential issue since most of Twitter’s revenue depends on advertisers.

Brands notably raised concerns about the company’s planned approach to content in 2022 in the lead up to Elon Musk acquiring Twitter. Musk responded to these concerns in an open letter in October 2022: “…Twitter obviously cannot become a free-for-all hellscape, where anything can be said with no consequences,” he wrote. He added, “Twitter aspires to be the most respected advertising platform in the world that strengthens your brand and grows your enterprise.” Many observers saw the hiring of Linda Yaccarino, a former media executive whose previous team generated more than $100B in ad sales, as Twitter CEO was a first step towards achieving this goal.

In a move that seemed to depart from this promise, Musk largely eliminated the T&S team and has been criticized for inconsistent moderation decisions since taking ownership of the company. In fact, Musk’s sharing of an antisemitic conspiracy theory in late 2023 led major brands including Disney, Warner Bros. Discovery, and Comcast to end their advertising. The same week, IBM stopped its advertising because their ads were being shown next to content about Nazis and Adolf Hitler. Researchers at progressive non-profit Media Matters wrote about several corporate ads being shown next to overtly antisemitic content.

By the end of November 2023, the X brand had plummeted with at least 100 advertisers stopping their advertising. The New York Times reported spend was down 60% in 2023 and that the company had experienced a potential loss of up to $75M in revenue.

To T&S professionals, this marked a major fall from the company’s previous positioning on safety. In 2016, Twitter, a household name, set out to establish itself as “the place to go” for news, moving on from being a social network to connect with friends and family. But with more growth, came more scrutiny. Between 2015 and 2016, the misogynistic harassment campaign GamerGate compelled the company to address doxxing, rape and death threats targeting women. A few years later, the company spoke up again in 2017 after confirmation of Russian interference in the 2016 U.S. elections, saying they wished they had done more to prevent abuse on the platform.

These events were each followed by significant investments in T&S in the form of back-end enforcement and front-end UX features and design. The company formed a public facing Trust and Safety Council in 2016 to engage safety advocates, researchers, and community groups covering a wide range of issues. Twitter also launched innovative safety product features at the time that gave people more control over their feed such as mute and blocking, Birdwatch/Community Notes which gave people more context about borderline content, and Rethink, a prompt which reduced future offending tweets. In 2018, Twitter acquired Smyte, which added data, product and engineering capacity. By this time, the company was telling civil society that the majority of enforcement was now proactive, illustrating a significant shift.
These investments paralleled the platform’s user growth from 2017–2022, when monetizable daily active users grew 118%. However, since Musk’s acquisition of the platform and subsequent divestment of T&S, X lost approximately 13% of daily users from 2022–2023. Lower T&S standards have undoubtedly led both users and advertisers to walk away – leading to its significant decline in revenue.
## Expert Investors

### T&S Investors with Operator Experience

<table>
<thead>
<tr>
<th>Investor</th>
<th>Representing</th>
<th>T&amp;S Investment</th>
<th>Operator Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dror Nahumi</td>
<td>Norwest Venture Partners</td>
<td>ActiveFence</td>
<td>ECI, Axonlink, AT&amp;T Bell Labs</td>
</tr>
<tr>
<td>Izhar Armony</td>
<td>CRV</td>
<td>ActiveFence</td>
<td>Onyx Interactive</td>
</tr>
<tr>
<td>Sara Ittelson</td>
<td>Accel</td>
<td>Cinder</td>
<td>Faire, Uber</td>
</tr>
<tr>
<td>Christine Keung</td>
<td>J2 Ventures</td>
<td>Cinder</td>
<td>Box</td>
</tr>
<tr>
<td>Emil Michael</td>
<td>Emil Michael</td>
<td>Checkstep</td>
<td>Uber, Klout</td>
</tr>
<tr>
<td>Jonathan Oringer</td>
<td>Jonathan Oringer</td>
<td>Checkstep</td>
<td>Shutterstock</td>
</tr>
<tr>
<td>Charles Songhurst</td>
<td>Charles Songhurst</td>
<td>Checkstep</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Lauren Wagner</td>
<td>Lauren Wagner</td>
<td>Cove</td>
<td>Meta, Google</td>
</tr>
<tr>
<td>Daniel Rumennik</td>
<td>Daniel Rumenik</td>
<td>Hive</td>
<td>Airbnb</td>
</tr>
<tr>
<td>Vic Singh</td>
<td>Vic Singh</td>
<td>Hive</td>
<td>Kansas Labs</td>
</tr>
<tr>
<td>Josh Hannah</td>
<td>Matrix Partners</td>
<td>Inflection Risk Solutions</td>
<td>Betfair.com</td>
</tr>
<tr>
<td>Mika Salmi</td>
<td>Lakestar</td>
<td>Modulate</td>
<td>AtomEntertainment, Viacom, Mad Races, Pressplane, CreativeLive</td>
</tr>
<tr>
<td>Michal Simkovic</td>
<td>Michal Simkovic</td>
<td>Pex</td>
<td>Circle (Local Social App), Around by Miro Labs</td>
</tr>
<tr>
<td>Meka Asonye</td>
<td>First Round Capital</td>
<td>SafetyKit</td>
<td>Stripe, Mixpanel</td>
</tr>
<tr>
<td>Alison Lange Engel</td>
<td>Greycroft</td>
<td>Spectrum Labs</td>
<td>Stripe, LinkedIn</td>
</tr>
<tr>
<td>Steve Vassallo</td>
<td>Foundation Capital</td>
<td>Trust Lab</td>
<td>Ning</td>
</tr>
<tr>
<td>Dafina Toncheva</td>
<td>US Venture Partners</td>
<td>Trust Lab</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Lance Priebe</td>
<td>Lance Priebe</td>
<td>Two Hat</td>
<td>Hyper Hippo</td>
</tr>
</tbody>
</table>
Appendix I

Data Sources and References

For more information on the T&S industry and its impact on the business, please see the below resources.

A. Paladin’s Safety Tech Reports
   a. [Paladin’s International State of Safety Tech 2023 Report](#): While this Report focuses on platforms and the trust & safety solutions they can deploy, Paladin’s report assesses the global Safety Tech landscape which also includes law enforcement, civil society, device-based protection, network filtering, and physical safety tech. Please see their report and the other research they have published for further understanding of this sector.
   b. [U.S. Safety Tech Report](#)
   c. [U.K. Safety Tech Report](#)

B. FactMR’s [Global Content Moderation Solutions Market Report](#)

C. Transparency Initiatives
   a. The Action Coalition on Meaningful Transparency’s [Transparency Initiatives Portal](#) is a central hub for tracking transparency initiatives, outputs, and tech policy events.
   b. The Sustainability Accounting Standards Board’s (SASB) [research project](#) for content governance in the internet media & services industry, dedicated to building transparency reporting metrics for investors.

D. More learning on Trust & Safety
   a. [TSPA T&S Curriculum](#)
   b. [DTSP’s Glossary of Trust & Safety Terms, Best Practices Framework](#)
   c. [All Tech is Human T&S Knowledge Hub](#)
   d. [Data & Society Origins of T&S](#)
   e. [T&S Teaching Consortium](#)
   f. Integrity Institute’s [Best Practices for Start Ups and Early Stage Companies](#)
Appendix J

Acknowledgments / Contributors

Financial support for this report was provided by the Future of Online Trust and Safety Fund, a fiscally-sponsored multi-donor fund at Global Impact that supports charitable activities to build a more robust, capable, and inclusive Trust & Safety ecosystem and field.

Authors

- Joe Ammirato, Associate, Duco
- Vivian Chong, Senior Fellow, Duco
- Josh Lawson, Director, AI & Democracy, Aspen Institute
- Maura O’Sullivan, Expert, Duco
- Sarah Oh, Non-Resident Fellow, DFR Lab Atlantic Council
- Lauren Wagner, Fellow, Berggruen Institute

Thank you to Morgan Felchner, who assisted with the editing of the report, and JooHo Yeo, who supported the analysis of the report.

Contributors

Thank you to the following individuals for their contributions to this report:

- Sofia Arimany, Director of Tech Policy, Duco
- Connie Chung, Head of Policy, Whatnot
- Kat Duffy, CEO, RightsDuff Strategies
- Michael Dworsky, CEO & Co-founder, Cove
- Brian Fishman, CSO & Co-founder, Cinder
- Kelly Graziaedi, General Partner & Co-founder, f7 Ventures
- Katie Harbath, Chief Global Affairs Officer, Duco
- Lucia Stacey Harris, Head of T&S Solutions, Cinder
- Sara Ittelson, Partner, Accel
- Vaishnavi J, Expert, Duco
- Agne Kaarlep, Head of Policy & Advisory, Tremau
- Christine Keung, Partner, J2 Ventures
- Sidney Olinyk, CEO & Founder, Duco
- John Redgrave, VP T&S, Discord
- Aaron Rodericks, T&S Consultant
- Alex Rosenblatt, Founder, SafetyKit
- Yoel Roth, Knight Visiting Scholar, University of Pennsylvania
- Micah Schaffer, T&S Consultant
- Tom Siegel, CEO & Co-founder, TrustLab
- Derek Slater, Co-founder, Proteus Strategies
- Dave Willner, T&S Consultant
- Glen Wise, CEO & Co-founder, Cinder
- Shu Dar Yao, Managing Partner & Founder, Lucid Capitalism
- JooHo Yeo, MBA Candidate, Stanford GSB