

# Advancing Trust & Safety

Systems and standards for online safety professionals



# About 5Rights Foundation

5Rights develops new policy, creates innovative frameworks, develops technical standards, publishes research, challenges received narratives and ensures that children's rights and needs are recognised and prioritised in the digital world. While 5Rights works exclusively on behalf of and with children and young people under 18, our solutions and strategies are relevant to many other communities.

Our focus is on implementable change and our work is cited and used widely around the world. We work with governments, intergovernmental institutions, professional associations, academics, businesses, and children, so that digital products and services can impact positively on the lived experiences of young people.

## About the author

Alexandra Evans has worked as a content regulator (British Board of Film Classification), at a non-profit (5Rights Foundation) and for a global tech company (TikTok). She is interested in understanding, and responding to, the impact of technology on society. Previously, Alexandra was a lawyer (Mishcon de Reya) specialising in public and human rights law.

MARCH 2025

# Contents

<b>Foreword</b> .....	4
<b>Introduction</b> .....	6
Key findings .....	7
Recommendations .....	8
<b>1. Overview of Trust &amp; Safety</b> .....	11
1.1 What is T&S? .....	13
1.2 A brief history of T&S .....	13
1.3 What does T&S do? .....	17
1.4 Who works in T&S? .....	22
1.5 Summary .....	23
<b>2. Challenges</b> .....	24
2.1 Business model .....	26
2.2 Having to make the case for safety .....	30
2.3 Reporting lines .....	35
2.4 Metrics .....	37
2.5 Safety by design .....	38
2.6 Resourcing and capacity .....	41
2.7 Working environment .....	43
2.8 Conflict of interest .....	49
2.9 Nascency of the profession .....	50
<b>3. Who can make change?</b> .....	54
3.1 Policy makers .....	56
3.2 Regulators .....	57
3.3 Tech companies .....	58
3.4 T&S organisations .....	59
3.5 T&S professionals .....	60
<b>4. Resources &amp; further reading</b> .....	62
<b>5. T&amp;S Assessment tool</b> .....	64
<b>Endnotes</b> .....	69

# Foreword

This report is intended to put Trust & Safety regimes on the radar as a tool to improve the efficacy of existing safety strategies and interventions. Agreed protocols and professional standards would give Trust & Safety teams authority to act and make clear that executives are responsible for the decisions they make or indeed fail to make. Trust & Safety cannot depend on good people alone. Good people need to operate within a good system. The conflict of interest between commercial and safety needs requires Trust & Safety professionals to be formally empowered and required to act in the best interests of society, citizens, and users, especially children.

Trade-offs of innovation and safety, of access and safety, of truth and safety, of crime and safety, of profit and safety — must reflect the extraordinary power of the tech sector to shape business, culture and society. Safety needs to be established in law, in regulation, and in professional standards. Only then can Trust & Safety teams operate in good faith. Allowing a single CEO focused on shareholder interests to be the sole or determining voice of a company's safety standards is not adequate, nor is the lack of safeguarding of those who have to work at the cutting edge of human depravity. Trust & Safety professionals need the authority to act independently in accordance with clear and understood expectations backed up by professional standards and regulatory oversight.

Already, Trust & Safety is a global profession of well over a hundred thousand people, but unless and until it operates according to understood and enforced standards, looks after its own, and fulfils its purpose of keeping those that engage with tech products and services safe, it remains part of the problem not the solution.

I am grateful to all those who contributed to this report, particularly those who have worked in Trust & Safety teams for giving their expertise so generously. My particular thanks to Leanda Barrington-Leach, Arturo Béjar, Dr Richard Graham, Toby Shulruff, Matthew Soeth, and Vaishnavi J. I thank the author, Alexandra Evans, for this thoughtful piece of work, and trust that those responsible for the safety of citizens, children, and customers, will find something in its pages to help the sector live up to its title.



BARONESS BEEBAN KIDRON  
*5Rights Founder & Chair*

---

**We don't just need  
to be hired,  
we need to be powerful.**

SAHAR MASSACHI, INTEGRITY INSTITUTE, APRIL 2024

---

# Introduction

Trust & Safety (T&S) professionals play a key role in creating safer digital spaces and yet their role is poorly understood and therefore under considered. Based on the testimony and commentary of current and former T&S professionals (including whistleblowers), guidance and reports published by T&S membership bodies, and academic research, this report describes what T&S is, what T&S professionals do, and the barriers they face.

The report is for policy makers, politicians, and regulators setting and enforcing minimum standards for online safety. It will inform the work of civil society and academics as they seek solutions to make the digital environment safer. It is an opportunity for tech companies to evaluate their own systems and processes, to address areas where they fall short in enabling T&S's work, and to increase transparency including sharing best practice. Finally, the report aims to inform, support, and celebrate those working in this nascent profession and to encourage them to push for change.

# Key findings






- The systems and processes in which T&S professionals operate are optimised for profit rather than safety. This undermines T&S's central duty to make digital products and services safe for citizens, society and users.
- In the absence of a clear mandate, T&S professionals report having to make the case for safety standards repeatedly and facing significant challenges in fulfilling their brief.
- The combination of the importance of their day-to-day work and the systemic challenges they face takes its toll on many T&S professionals' mental health and wellbeing.
- Good people are not a substitute for good systems.
- If the systems and processes in which T&S professionals operated were fit for purpose, digital services and products would be safer.

# Recommendations

In response to these findings, we propose a series of recommendations to establish T&S as an independent, fully empowered profession working within or on behalf of tech companies, but in the best interests of society, citizens, and users – especially children.

Some of these recommendations may already be in place in some organisations, but the evidence suggests coverage is partial. The recommendations are a starting point for further discussion among T&S professionals and other key stakeholders.

## Give T&S professionals a clear mandate

-  Codify the right and responsibility of T&S professionals to act in the best interests of society, citizens, and users – even when this duty conflicts with the business interests of the service provider.
-  Ensure T&S professionals have the authority to proactively surface risk and to implement safety initiatives across all aspects of the service and product design and delivery through professional standards and regulation.
-  Embed T&S into all stages of design and development processes, granting T&S professionals a mandate to modify, delay, halt, or withdraw services, products, AI systems (including algorithms and generative AI), features, and functionalities if safety risks have not been sufficiently mitigated.
-  Include meaningful and consistent safety metrics in company-wide targets with the same status as other business goals.
-  Increase transparency about the systems and processes in which T&S professionals operate including sharing best practices.



## Ensure governance systems and protections are sufficient

- Create a direct reporting line from T&S leadership to the company's CEO/senior leader.
- Require companies to keep written records of advice from T&S to senior leadership, including when leadership ignores warnings or rejects recommendations.
- Enhance support and protections for those who raise safety concerns, including by codifying the four principles of the Right to Warn<sup>2</sup> across the entire tech sector.
- Provide differentiated — but equivalent — remuneration and incentivisation schemes to protect the independence of T&S professionals.

## Create safe and sustainable working practices

- Develop, and consistently apply, industry-wide minimum health and safety standards to protect the physical and mental health of all T&S professionals — irrespective of geography or contract type.
- Allocate appropriate resources to safety teams, based on T&S leadership's assessment of needs and safe operating capacity.
- Establish enhanced procedures when tech companies make staffing cuts that may impact safety, including mandatory risk assessments and a requirement to notify relevant regulatory authorities.

## Establish professional standards, protections and training

- Clarify the skills and qualifications required for different roles within T&S and create validated training programmes to support new appointments and continued professional development.
- Expand the remit of T&S membership organisations to include representing the needs, interests, and concerns of T&S professionals to policy makers, regulators, and tech companies.

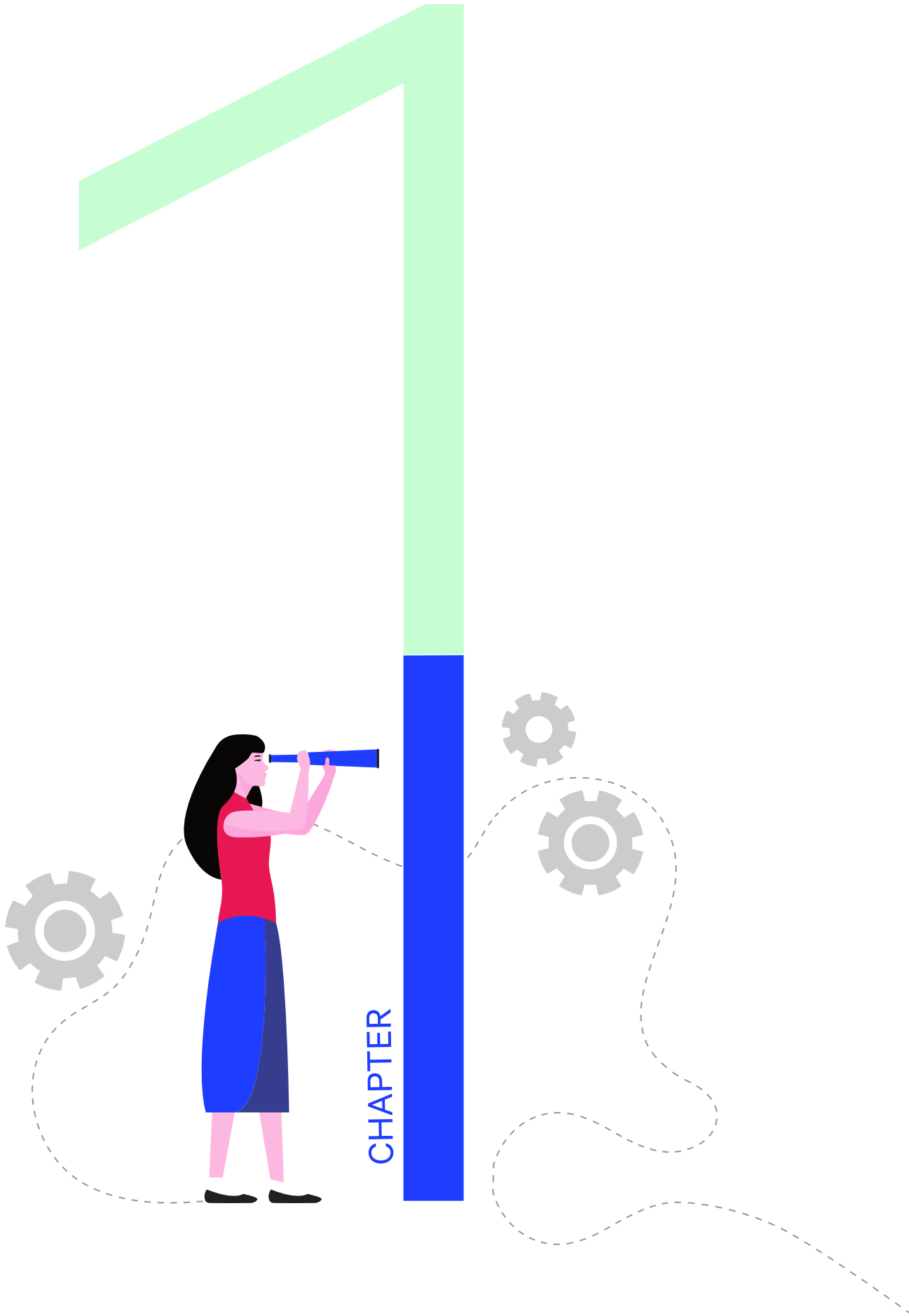
- Develop professional codes of practice to guide and protect those working across the T&S ecosystem.

## Beyond T&S

Addressing the structural and governance challenges faced by T&S is important and urgent work, but the changes required to create a digital environment fit for childhood go far beyond enhancing T&S's status and ensuring its independence. Governments must:

- Establish comprehensive and legally enforceable minimum standards that ensure digital products and services are age-appropriate, safe, private, secure, and rights-respecting by design and by default.
- Fully resource and empower regulators to hold tech companies accountable when they fall short.
- Increase accountability for tech CEOs and other executive leaders who are reckless about safety and the negative impacts of their products and services on society, citizens, and users.
- Require tech companies to give independent, accredited researchers access to their data so they can study online harms and identify mitigations.
- Require tech companies to remove persuasive design strategies that compromise personal freedom and agency, so users – especially children – can choose how, when, and for how long they engage with digital products and services.
- Include generative AI systems in policy and legislation.

Creating digital services and products that serve the interests of society, citizens, and users requires a fundamental shift in the accountability of tech companies. Skilled, accountable, and empowered T&S teams are central to the success of a responsible tech sector.



CHAPTER

---

**Trust & Safety (T&S)  
teams are most often  
born in a crisis...  
You grab whoever  
you can to address  
the problem immediately,  
and that's where T&S  
teams come from.**

KAREN MAXIM, JOSH PARECKI AND CHANEL CORNETT,  
T&S LEADS AT ZOOM<sup>3</sup>

---

# 1. Overview of Trust & Safety

## 1.1 What is T&S?

Trust & Safety (T&S) refers to those whose job it is to ensure the safety of digital products and services, including upholding Community Rules or Guidelines and responding when harms occur. Sometimes known as Integrity workers,<sup>4</sup> there are over 100,000 T&S professionals globally.<sup>5</sup> The harms they address include child sexual exploitation and abuse, exposure of minors to age-inappropriate themes, terrorism, violence, pornography, human trafficking, discrimination and hate speech, bullying and harassment, mis- and disinformation, suicide and self-harm, eating disorder and body image content, dangerous challenges, sale or promotion of illegal goods and services, drug and alcohol misuse, fraud and scams, and mental health and wellbeing issues, including compulsive use of technology.

## 1.2 A brief history of T&S

The internet was founded on the principle that it should be a free and open space beyond the constraints of 'real world' institutions and governance structures. Early pioneers championed unfettered free speech, protected by a shield of anonymity as a way of democratising access to public discourse and bypassing traditional gatekeepers.

The early online communities that formed in group chats, messaging and bulletin boards, forums, blogspheres, virtual worlds, and social networks broadly subscribed to these values and were generally self-governing. In practice, this meant rules were kept to a minimum, decisions about anti-social behaviour were reached collectively or by members nominated by users to make decisions on the community's behalf. Additionally, user controls such as blocking appear to have been favoured over content removal as a mechanism for managing exposure to harm or resolving conflicts.

John Perry Barlow was a Founding Member of the Electronic Freedom Foundation. On 8 February 1996, Barlow published his “Declaration of the Independence of Cyberspace” which captures the ideological perspective of the internet’s early founders and pioneers.

It begins:

### **Declaration of the Independence of Cyberspace**

*“Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask you of the past to leave us alone. You are not welcome among us. You have no sovereignty where we gather.*

*We have no elected government, nor are we likely to have one, so I address you with no greater authority than that with which liberty itself always speaks. I declare the global social space we are building to be naturally independent of the tyrannies you seek to impose on us. You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear.*

*Governments derive their just powers from the consent of the governed. You have neither solicited nor received ours. We did not invite you. You do not know us, nor do you know our world. Cyberspace does not lie within your borders. Do not think that you can build it, as though it were a public construction project. You cannot. It is an act of nature and it grows itself through our collective actions...”<sup>6</sup>*

However, even in the early days, when online communities were still relatively small and homogenous, agreeing rules and responding to abuse could present challenges. The case of LambdaMOO and Mr Bungle, as reported in 1993 in “A Rape in Cyberspace”,<sup>7</sup> demonstrates that questions about what is and isn’t acceptable online, who is responsible for oversight of digital spaces, and how to respond are not new.

In 1995, a landmark US judgement<sup>8</sup> held that online spaces could be subject to ‘real-world’ laws. Prodigy, an online forum that billed itself as a safer, more moderated space was successfully sued by Stratton Oakmont (the brokerage firm featured in the *Wolf of Wall Street*) for defamation over an anonymous post on its “Money Talk” board. The judge held that Prodigy’s interventionist approach to moderation meant it had assumed editorial control.

In 1996, the US introduced s.230 of the Communications Act<sup>9</sup> giving online service providers immunity from liability for content published by third parties and the right to restrict access to content.

---

**“No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”**

S.230(C)(1) COMMUNICATIONS ACT 1934<sup>10</sup>

---

Whilst s.230 immunity is not absolute, it has been determinative in shaping the digital environment in its current form including the way large companies approach safety.<sup>11</sup>

---

**“So many of these very dominant platforms were built [in the US], right, by founders who grew up in an environment where the First Amendment and this notion of a marketplace of ideas and a bunch of the things that come with that is built into their product and their vision for their product.”**

NICOLE WONG, FORMER VP AND DEPUTY GENERAL COUNSEL  
AT GOOGLE AND FORMER LEGAL DIRECTOR OF PRODUCTS AT TWITTER<sup>12</sup>

---

In the 2000s, as wifi, broadband, and 3G rolled out, the internet began to take shape in the form we recognise today. With s.230 as a sword and a shield, a small group of companies established market dominance, and their products and services became ubiquitous.<sup>13</sup> Since then, the United State’s safe harbour protections have given a handful of tech CEOs unprecedented power to determine content and conduct standards online. Their decisions impact the safety, privacy, and security of billions of people worldwide.

---

**“In a lot of ways Facebook is more like a government than a traditional company.”**

MARK ZUCKERBERG<sup>14</sup>

---

The consolidation, centralisation, and commodification of digital spaces as well as the rapid increase in user numbers brought heightened awareness of the potential for harm, especially for minority groups and children. As platforms grew, the range and scale of safety issues also expanded.

These risks were addressed reactively and iteratively<sup>15</sup> typically by existing employees working in other teams tasked with solving specific challenges as they arose. Soon those with an interest in solving safety challenges coalesced and then emerged as a distinct entity within tech companies. The term “Trust & Safety” was coined by Rob Chesnut, Jeff Taylor and Lisa Laursen at eBay in 2002.<sup>16</sup>

---

**“To our knowledge, it was the first Internet ‘Trust & Safety’ department, so named because it focused on building trust among strangers and kept our site safe for the world to use.”**

ROB CHESNUT, FORMER SENIOR VICE-PRESIDENT, TRUST & SAFETY, EBAY<sup>17</sup>

---

As public concerns about safety grew, and to ward off threats of oversight, tech companies began to propose self-regulatory and voluntary initiatives. These focused on greater cross-industry cooperation,<sup>18</sup> promoting digital literacy,<sup>19</sup> development of parental controls,<sup>20</sup> improved content detection strategies,<sup>21</sup> and enhanced transparency.<sup>22</sup> T&S teams had responsibility for the developing and implementing these strategies.

Within tech companies, T&S teams appear to have proactively sought to understand and prevent risk of harm, often working with NGOs and academics<sup>23</sup> or recruiting such experts into safety roles.<sup>24</sup> At the same time, safety campaigners in civil society argued for a comprehensive response, emphasising a much greater focus on prevention and safety by design (see below).

Frustrated by the failure of self-regulation and the reluctance of tech companies to address harms, lawmakers in many countries and regions have now begun to intervene to set minimum standards and to appoint regulators to oversee compliance.<sup>25</sup>



---

“Trust & Safety (T&S) teams are most often born in a crisis..You grab whoever you can to address the problem immediately, and that’s where T&S teams come from.”

KAREN MAXIM, JOSH PARECKI AND CHANEL CORNETT, T&S LEADS AT ZOOM<sup>26</sup>

---

In many countries, regions and states, companies are now required to meet legal obligations on issues such as risk assessment, recommender systems, safety by design, and age-appropriate design principles. This is likely to mean that T&S’s role has a greater focus on compliance and their priorities and approach likely to be increasingly shaped by legal teams.

In the future, the remit of T&S will continue to evolve. For example, advances in AI will further reduce the role carried out by human moderators. Generative AI also creates new risks that providers of both traditional and emerging digital services and products must anticipate and respond to.

### 1.3 What does T&S do?

The greatest number of T&S professionals are content moderators.

#### What is a content moderator (CM)?

*“CMs enforce the online rules which tell users how to behave and what content is acceptable on a particular internet site (Gerrard, 2022). Some sites use a “community reliant approach” where communities define their own standards that are then put into effect by volunteer moderators, whilst other sites use an industrial approach, where workers are employed to enforce a set of standardised rules (Caplan, 2018). As such, the ways in which CMs are employed vary.*

*Whilst some are volunteers, paid moderators can range from in-house workers employed directly by the company needing moderation; to boutique firms specialising in content moderation for other companies;*

*to outsourced third party vendors and microlabour platforms (Roberts, 2019). This means CMs can experience a range of working conditions where some have less workplace protection as technically, they are not employees - thus the company requiring moderation work has a level of “plausible deniability” to the harm faced by these CMs (Barrett, 2020; Roberts, 2019).”*

THE PSYCHOLOGICAL IMPACTS OF CONTENT MODERATION ON CONTENT MODERATORS: A QUALITATIVE STUDY, 2023<sup>27</sup>

Within content moderation, a clear division has emerged: those employed by tech companies in their larger safety hubs enjoy far better pay, benefits, and conditions. In contrast, those working in the Global South or via sub-contractors are likely to have poorer working conditions and pay.<sup>28</sup>

Whilst content moderation is an important aspect of T&S work, the field encompasses much more. All online content moderators work in T&S, but not all those who work in T&S are content moderators. This report considers the experiences and challenges of content moderators, including the inequities and challenging working conditions they experience.<sup>29</sup> However, when considering the role of T&S professionals in shaping online safety standards, it is necessary to distinguish frontline moderators from the much smaller and arguably more powerful group of ‘HQ’ T&S professionals.

This group plays a crucial role in defining and implementing safety standards. Their responsibilities include writing platform policies (both front-facing policies such as Community Guidelines and the more detailed policies that content moderators apply in their day-to-day work); adjudicating ‘edge case’ content and issues (escalations) that require specialist review by senior experts; developing and deploying the machines that proactively detect potentially violative content, contact, and conduct; responding to requests from and/or reporting incidents to national and international law enforcement and security services; resourcing and managing the work of content moderation teams; collecting and analysing data relevant to safety (metrics); developing safety tools and resources (e.g. parental controls or safety hubs); engaging with external experts (e.g. academics and NGOs); advising on the levels of safety of new product and service proposals; supporting compliance; and transparency reporting.<sup>30</sup>

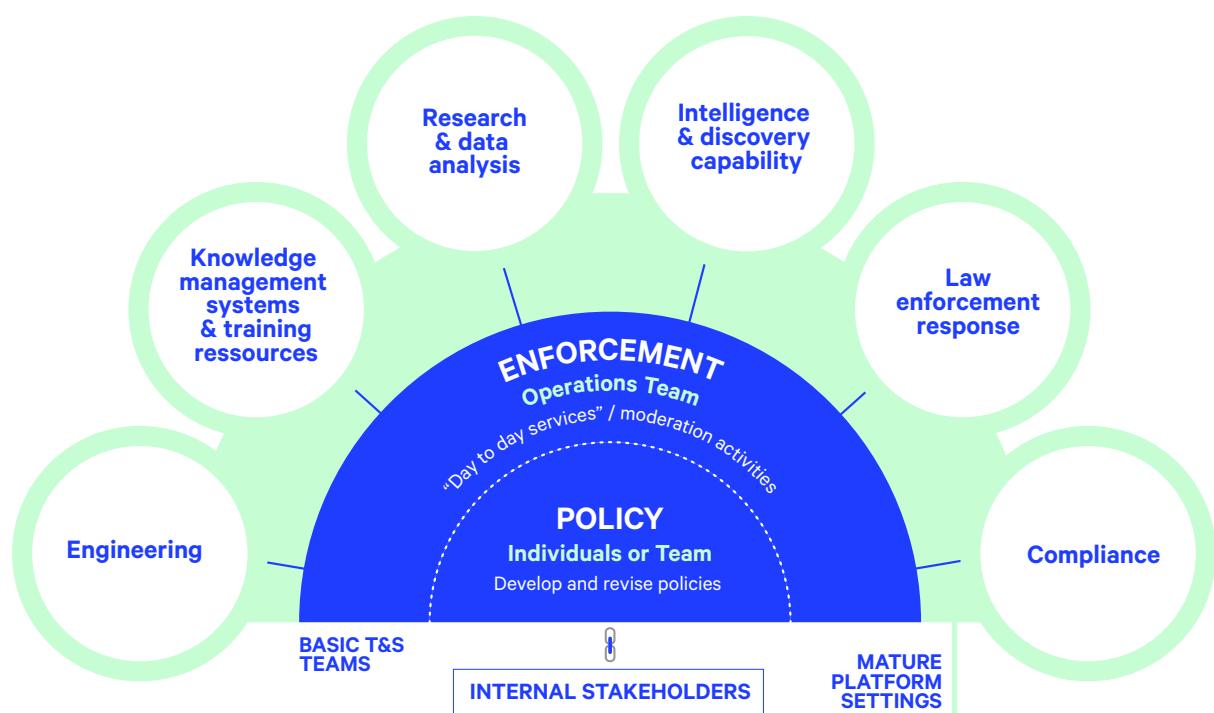
The majority of quotes from T&S professionals in this report come from those working in ‘HQ’ T&S roles because they are in a position to shape the safety strategy of digital

services and products. By contrast, whilst content moderators play a critical role in protecting users, they have very limited agency or influence over a tech company’s safety strategy (in itself a cause of the mental health and wellbeing challenges many content moderators experience).

## Key elements of a T&S team

According to the Trust & Safety Professional Association’s (TSPA)<sup>31</sup> Introduction to Trust & Safety<sup>32</sup> resource, the basic components of most T&S teams are the policy and enforcement teams. The policy team sets “rules of the road” (often referred to as Community Rules or Guidelines). The enforcement team enforces them typically through a combination of automated and human moderation.

Platforms with larger, more mature T&S functions may also include engineering teams that build and maintain moderation tools, training and knowledge management for moderators, researchers and data analysts who provide the evidence base that informs all aspects of T&S work including the level and prevalence of harm, and intelligence and discovery teams to detect emerging threats. The law enforcement response and compliance team handle legal requests and ensure regulatory compliance (often in collaboration with Legal).



How T&S teams are structured and how they collaborate and provide input into the work of other teams varies from company to company. TSPA describes two main approaches to structuring T&S teams.

*“A **centralized** Trust & Safety model consists of a single, discrete team that is responsible for all Trust & Safety components from end-to-end. This single team works hand-in-hand with product teams but is ultimately tasked with owning policy creation and enforcement, as well as advising at the product ideation and development stage to ensure safety is incorporated into the product, a concept known as Safety by Design.”*

*“A **dispersed** Trust & Safety approach is a model in which Trust & Safety team members are distributed throughout a company, and there is no single, central team. Rather, Trust & Safety professionals are embedded throughout the company and report to the head of each business unit, rather than a singular head of Trust & Safety.”<sup>33</sup>*

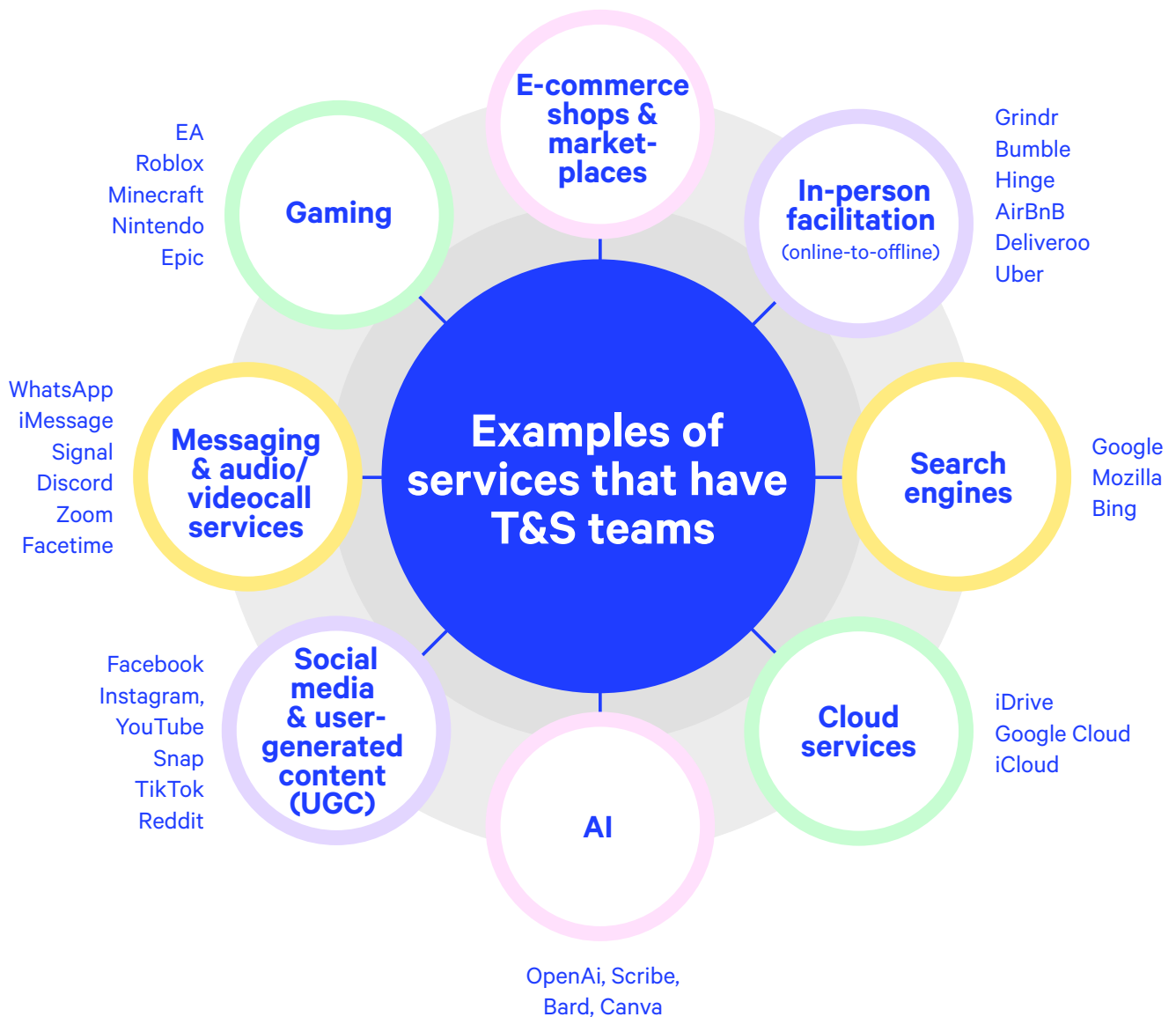
The longer a tech company has had a T&S team, the more likely it is to adopt a centralised model.<sup>34</sup>

How a company determines its Trust & Safety needs and the way it organises the function as a result depends on numerous factors, such as:

- The type of product or service (e.g. social media network, e-commerce, search engine, gaming, messaging, video conferencing or live streaming);
- The types of abuse, misuse, and disruptive conduct the company must address;
- The set of values that the company upholds;
- The demographics of its customer base;
- The countries in which it operates;
- The size and maturity level of the company.<sup>35</sup>

TSPA’s explanation of the main operating models suggests that in some companies, some of these functions may be partially integrated into T&S or vice versa. In small or low-risk services, there may be no dedicated T&S function, whilst the largest services have tens of thousands of T&S workers.

Whilst they may vary in size and structure, it is typical for major platforms to have a T&S function. The illustration below provides examples of the types of digital products and services that have employ T&S professionals.



In 2022, Snap increased its Trust & Safety budget to “approximately \$164 million”, though it subsequently decreased spending on Trust & Safety issues to \$135 million in 2023.<sup>37</sup> In the same year, Amazon reported investing more than \$1.2 billion and employing more than 15,000 people to address counterfeit, fraud, and other forms of abuse.<sup>38</sup> TikTok said it would spend US\$2bn in 2024 on safety.<sup>39</sup> Whilst it is hard to verify claims by tech companies about how much they spend on safety, these figures give some indication of the scale of T&S operations.

Although its origins may be modest, Trust & Safety is now a multi-billion-dollar industry, encompassing a number of third-party providers including Accenture, Concentrix, GenPact, TaskUs, and Teleperformance.<sup>40</sup> These companies are part of a wider safety tech sector which is growing at pace.<sup>41</sup>

## 1.4 Who works in T&S?

Because T&S is a relatively new field, many people working in the sector started out in different careers. For example, they may be former academics,<sup>42</sup> law enforcement officers, data scientists, journalists, or have worked in government,<sup>43</sup> education, the military, or for an NGO or civil society organisation.<sup>44</sup>

When considering what motivates people to choose a career in T&S, job descriptions and accounts from T&S professionals highlight a combination of intellectual challenge, pace, complexity, and the opportunity to have impact.<sup>45</sup>

### T&S skills and mindset

Research carried out by Toby Shulruff at Arizona State University found that four key areas of skill and experience are important for T&S work. These are:

- Analysis, critical thinking, or research
- Project management and related skills
- Investigation
- Subject Matter Expertise in abuse areas

In addition, eight mindsets important for T&S work were identified:

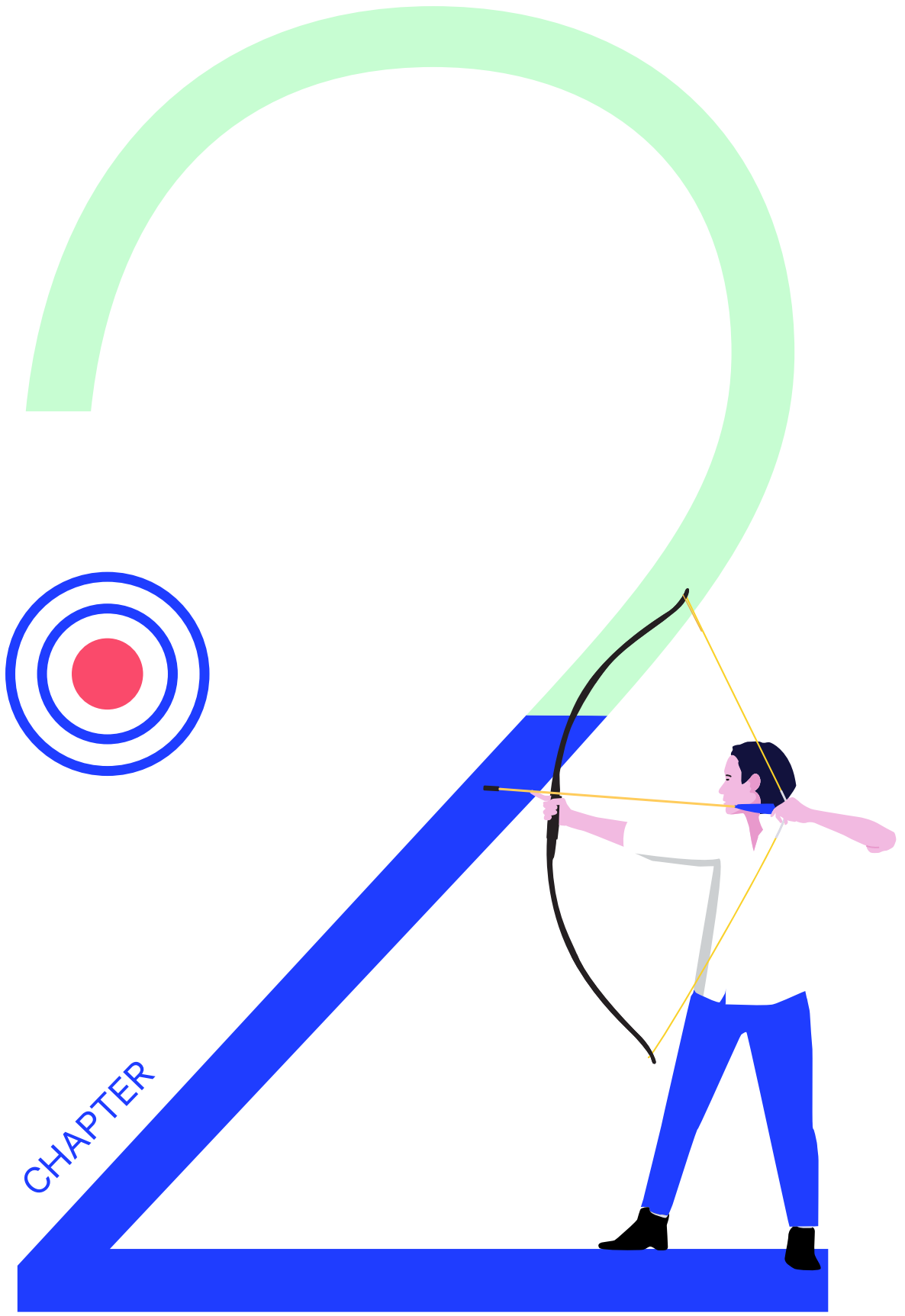
- Problem-solving
- Communication
- Passion
- Collaboration
- Adaptability or flexibility
- Curiosity
- Empathy and resiliency
- Ability to understand complexity or context

T&S professionals frequently describe a sense of mission and cite the fact that their work is societally important as a key motivator for choosing their career.

The rapid growth of tech companies also often provides opportunities for rapid professional growth and the chance to assume responsibility early on. For some, working in T&S is not only interesting but also well paid. The average salary for a T&S Safety Analyst in the US is US\$62,955,<sup>46</sup> while managers can earn up to US\$81,274.<sup>47</sup> Senior T&S professionals at large tech firms earn significantly more.<sup>48</sup>

## 1.5 Summary

As the review of its history shows, T&S has emerged rapidly over the last two decades. Teams are born from crisis and take shape iteratively. Given the importance of T&S's role in making digital products and services safer, the nascency of the profession, the range of platforms that have T&S teams, the number of people who do T&S work globally and the disparities in their working conditions and wages, there is an urgent need to establish organising principles for the professional. Without these, T&S professionals face significant barriers when protecting society, citizens and users. These are explored in Chapter Two.





---

**It was growth  
at all costs, and safety  
eventually.**

DEL HARVEY, FORMER HEAD OF TRUST & SAFETY, TWITTER<sup>49</sup>

---

# 2. Challenges

Through whistleblower testimony, academic research, and reports and guidance published by T&S professionals and emerging professional bodies, we have begun to gain a better understanding of the role T&S plays and the challenges T&S professionals face.

It is wrong to suggest that T&S is ineffective. The policies, processes, and technologies developed and deployed by T&S play an essential role in managing online risk including emergency responses during moments of crisis. If T&S professionals were to collectively down tools, the consequences for the safety, privacy, and security of online spaces would be immediate and severe. Equally, the recent layoffs across T&S have prompted stakeholders to raise concerns that platforms may be less safe as a consequence (see Resourcing and capacity below). However, **evidence from within the profession suggests that T&S professionals experience significant structural and commercial barriers that hinder their ability to operate effectively.**

## 2.1 Business model

—

“In working for a private, for-profit company, your job is ultimately tied to helping that company make money... As painful as it may be to acknowledge, go in with the expectation that you’re working under [an] incentive structure that if push came to shove, ultimately revolves around profit. Plan accordingly.”

LEADERSHIP ADVICE FOR NEW TRUST & SAFETY LEADERS, INTEGRITY INSTITUTE<sup>50</sup>

—

There is an inherent tension between the profit imperative of commercial digital platforms and T&S’s function which is to keep society, citizens, and users safe.

—

“Design strategies on social media are shaped by three broad commercial goals: to increase the number of users, to maximise the amount of time users spend on

the service, and to increase the amount of content generation and interaction with the service. As designers themselves acknowledge, ‘reducing attention will reduce revenue.’ These business objectives shape design strategies, and in turn, outcomes for children.”

UPDATED DISRUPTED CHILDHOOD: THE COST OF PERSUASIVE DESIGN, 5RIGHTS FOUNDATION<sup>51</sup>

---

## Attention Economy

A common business model for digital platforms is the ‘attention economy’ model where services exchange access for users’ attention and personal information (data) which they then monetise through ads revenue. Aspects of product and service design that support this model (e.g. deployment of persuasive design strategies, ultra personalised recommender systems and encouragements to expand social networks and engagement) can undermine user safety and make it hard to disengage — even when doing so is in a user’s best interests.<sup>52</sup>

Remedies that enhance safety such as identifying and removing underage users, age-restricting access to features and functionalities, limiting social connection recommendations or dispersing filter bubbles and rabbit holes have a direct impact on both user numbers and user engagement. This makes them unattractive to those with an interest in profitability and growth such as founders, investors, and shareholders.<sup>53</sup>

## Success metrics

---

**“It was growth at all costs, and safety eventually.”**

DEL HARVEY, FORMER HEAD OF TRUST & SAFETY, TWITTER <sup>54</sup>

---

User growth is a defining metric when measuring the value or success of a commercial online service. There is little commercial incentive to prioritise safety and clear disincentives to implementing safety initiatives that jeopardise user numbers or growth.

---

“I can say this as a founder, I am under tremendous pressure to deliver results... There’s, I think, a lot of pressure to do anything or say anything in order to succeed, and I think that can lead you down the road of doing things that are not good for or to users.”

THE UNSEEN TEEN, DATA & SOCIETY<sup>55</sup>

---

**It is therefore predictable that when safety strategies are considered, decision-makers responsible for the profitability of the company will favour solutions that do not negatively impact user numbers or other key metrics.** For example, they may opt for actions such as creating a safety centre, funding in-school digital literacy programmes, offering parental controls (which has minimal impact on the available pool),<sup>56</sup> and removing harmful content (which reduces the available content pool minimally).<sup>57</sup> These measures typically have little or no impact on user engagement or numbers but also do not address risks inherent in the business model and “baked in” to product design.

---

“I have spoken to many people working on Integrity or Trust & Safety at Meta and other companies. Often, they are demoralized... They feel that managers do not want them to reduce users’ exposure to unwanted and harmful content if it negatively impacts what’s called “engagement”. The companies’ main aim is to keep users clicking and scrolling, so they will spend more time on the service and see more advertising.”

ARTURO BÉJAR, INSTAGRAM WHISTLEBLOWER<sup>58</sup>

---

## Impact of safety on share price

In February 2022, Facebook announced its first-ever drop in daily active users (DAUs).<sup>59</sup> Meta’s share price fell by 25%, wiping US\$200bn<sup>60</sup> off the company’s value in a single day. In contrast, the day Frances Haugen testified before the Senate to highlight fallings in safety, Meta’s stock price rose albeit slightly.<sup>61</sup> One market analyst described Haugen’s revelations about safety as “a speedbump rather than a roadblock”.<sup>62</sup> Another noted “If the advertisers don’t leave, no one’s going to care and when we talk to advertisers, they say, ‘We don’t care’”.<sup>63</sup>

Whilst some market analysts argue that revelations about tech companies' safety record have a longer-term impact on user numbers and growth<sup>64</sup> as well as exposure to regulatory fines and litigation, they have not led to the kind of dramatic stock sell offs that slowdowns in user growth or sales can trigger.<sup>62</sup>


## Founder mindset

Another factor influencing safety decisions is the 'innovation mindset' of founders and leaders. **It is harder to argue for safety if leadership is sceptical that existing standards apply to their product strategy and instead believe that commercial success comes from moving fast and breaking things.**<sup>66</sup> For example, in 2010, Mark Zuckerberg addressed his company's decision to change the privacy settings of 350 million Facebook users, stating: *"A lot of companies would be trapped by the conventions... and we decided that these would be the social norms now and we just went for it."*<sup>64</sup>

The Federal Trade Commission disagreed with Zuckerberg's view that the digital revolution had rewritten societal norms on privacy.<sup>68</sup> However, the view that innovation takes place at the edge of — or outside — the boundaries of established standards and norms is fundamental to the business models of many tech companies. This mindset often frames safety (and those advocating for it) as a drag on progress and, by extension, profit.

Instead of adapting their business model or mindset, tech companies may choose to invest in defending them. For example, in 2019, former General Counsel for Apple, Bruce Sewell, revealed that Apple's legal budget when he left in 2017 was \$1 billion per year.<sup>69</sup> In an interview at Columbia Law School, Sewell also explained his job at Apple was not to stay clear of the line dividing legally risky actions from clearly safe actions, but rather to *"steer the ship as close to that line as you can, because that's where the competitive advantage lies ... you want to get to the point where you can use risk as a competitive advantage"*. A legal team becomes an asset when it helps the company to approach this blurry legal/illegal line strategically and then manage the "nuclear" situation, if trouble arises.<sup>70</sup>

This false tension between innovation and safety creates significant challenges for T&S professionals advocating for safety by design (see below).<sup>71</sup>

 **Codify the right and responsibility of T&S professionals to act in the best interests of society, citizens, and users – even when this duty conflicts with the business interests of the service provider.**

## 2.2 Having to make the case for safety

—

“The company may not devote the kind of resources to [safety] issues that they do to other areas in the company, even though, objectively, outsiders might consider them to be among the most important. Integrity, Trust & Safety workers face this kind of challenge every day.”

ARTURO BÉJAR, INSTAGRAM WHISTLEBLOWER<sup>72</sup>

—

In the absence of legal requirements (see below), senior leaders are free to deprioritise safety in favour of commercial goals (see above). As a result, T&S professionals must make the case for safety rather than being empowered to ensure tech companies set and follow minimum standards. This dynamic means that **safety proposals can be rejected, deprioritised or diluted in favour of other priorities, such as product strategy, share price, and advertising revenue and T&S lacks a clear mandate to object.**<sup>73</sup> Research published by the UK Government in 2024 revealed that T&S professionals working at large platforms consistently had to secure high levels of approvals for decision-making processes, such as escalation to Senior Directors, VPs, or C-Suites.<sup>74</sup>

### Negotiating with other teams

—

“As a Trust & Safety professional, your goals will inevitably collide with those of other departments... Product development teams want to roll out products as quickly as possible and can sometimes see building Trust & Safety features as non-essential work that delays launch.”

MAKING THE CASE FOR TRUST & SAFETY, SPECTRUM LABS<sup>75</sup>

—

It is over simplistic and unfair to suggest that only those working in T&S teams care about safety. As 5Rights Foundation's report *Pathways: How digital design puts children at risk*<sup>76</sup> makes clear, others (in this case product designers) within tech companies are also deeply concerned.

---

**“The designers interviewed were uncomfortable with the solely commercial intent of the companies they worked for but felt that change would only come if commercial goals specifically required them to design for the safety and wellbeing of children. Some acknowledged that the ‘products’ they were designing were bad for children, but they repeatedly expressed the need for change ‘from the top’.”**

PATHWAYS: HOW DIGITAL DESIGN PUTS CHILDREN AT RISK, 5RIGHTS FOUNDATION<sup>77</sup>

---

Whilst T&S does not have a monopoly on good intentions or moral clarity among tech employees, its unique remit within tech companies is to prioritise user safety. The fact that T&S teams must repeatedly make the case for safety – and can be overruled, particularly by those focused on the company's profitability – is a significant barrier to creating safe products and services by design and by default.

In October 2021, Spectrum Labs, a third-party content moderation provider, published its white paper, *Best Practices for Making a Trust & Safety Business Case*.<sup>78</sup> Drawing on its own experiences helping clients present the case for T&S, as well as interviews with T&S professionals, the report demonstrates that safety enhancements are secured by negotiation rather than by default. It also provides practical advice on how to persuade CEOs and other teams to prioritise safety including:

- *“Building a Trust & Safety business case can be a process, and you should be prepared for a long haul.”*
- *“Senior executives at emerging platforms and communities are understandably sensitive to barriers to growth... position Trust & Safety as a driver of growth.”*
- *“Learn and speak the language of each department.”*
- *“The reality we face today is that building a Trust & Safety business case may require a team of data scientists and analyst resources which you may not have access to.”*
- *“Lean into the moral argument and back it up with the impact of churn on the bottom line.”*

While there is nothing inherently wrong with advising T&S professionals on forging

better relationships with colleagues or persuading others to invest time and money in safety, this advice illustrates that there is no automatic mandate for T&S to insist that their work is prioritised.

**The level of safety provided by a platform should not depend on the advocacy skills or the ability of T&S professionals to convince colleagues that safety is a driver of growth (which it may not be). Neither should it rely on how receptive their CEO is to the “moral argument” for safety.**

## Central role of leadership

---

“My first CEO was engaged [in Trust & Safety], the second not at all – this makes a big internal difference.”

VP TRUST & SAFETY AT LARGE PLATFORM<sup>79</sup>

---

The T&S professionals cited throughout this report consistently describe the importance of senior decision makers in determining safety standards, especially CEOs. This has also been raised in legal action against major platforms. For example, the New York Attorney General’s lawsuit against TikTok includes the allegation that:

---

“TikTok employees have provided concrete suggestions for ways to make the platform safer, but those safety improvements were stymied by TikTok’s leadership’s pursuit of profits.”<sup>80</sup>

EXTRACT FROM THE LAWSUIT FILED AGAINST TIKTOK BY THE ATTORNEY GENERAL OF NEW YORK, OCTOBER 2024

---

Similarly, the New Mexico Attorney General’s lawsuit against Snap Inc (filed in October 2024) includes allegations from former Snap T&S employees that they were largely ignored by upper management and:

---

“that there was pushback in trying to add in-app safety mechanisms because [Snap CEO] Evan Spiegel prioritised design.”<sup>81</sup>

---



In January 2025, Meta announced changes to its safety policies and practices.<sup>82</sup> These changes included removing some restrictions on hate speech and abuse on the ground of sexual orientation, gender identity, and immigration status. Additionally, Meta will no longer use fact checkers and will stop proactively detecting and removing violative content for “lesser policy violations”. Instead, it will only review content when it is reported.

Meta has not confirmed which policies it considers “less severe”, but based on current removal rates, the impact on how much violative content is caught is likely to be significant. In Instagram’s transparency report for Q3 2024,<sup>83</sup> the percentage of violative content detected and actioned proactively by Meta compared to through user reports was (for example):

- bullying and harassment: 95.9% proactive vs 4.1% user reports<sup>84</sup>
- hate speech content: 98.5% proactive vs 1.5% user reports<sup>85</sup>
- violent and graphic content: 98.9% proactive vs 1.1% user reports<sup>86</sup>
- violence and incitement content: 99.3% proactive vs 0.7% user reports<sup>87</sup>

The New York Times reported that the decision to change the policies was made by Zuckerberg who chose a small number of senior employees including those from public policy and communications teams to consult with. “The entire process was highly unusual. Meta typically alters policies that govern its apps — which include Facebook, Instagram, WhatsApp and Threads — by inviting employees, civic leaders and others to weigh in. Any shifts generally take months. But Mr. Zuckerberg turned this latest effort into a closely held six-week sprint, blindsiding even employees on his policy and integrity teams.”<sup>88</sup>

---

**“I have a much greater command now of what I think the policy should be, and this is how it’s going to be going forward.”**

MARK ZUCKERBERG<sup>89</sup>

---

It appears that despite the significance of the change, typical policy approval processes were not followed. This highlights the challenge that T&S professionals face when safety decisions are subject to internal negotiations, with the CEO being the ultimate decision maker.

**Across all tech companies, greater transparency is needed on the extent to which T&S professionals have a mandate to raise objections**, how these objections are recorded and what happens in scenarios where they are asked to enact policy, system or design changes that they believe will make a product or service less safe — especially if they are not given sufficient opportunity to consider proposals or raise concerns.

It also raises the wider question of how T&S professionals raise and record objections and what happens when they are asked to enact policy, systems or design changes that they believe will make a product or service less safe especially if they are not given an opportunity to consider proposals or raise concerns.


## Creating a culture of challenge

The financial sector, another high-risk industry, illustrates how failures in risk management protocols can have severe consequences. Reviews into incidents<sup>90</sup> and guidance from regulators on how to manage risk<sup>91</sup> highlight the importance of creating a “culture of challenge” where those responsible for interrogating risk feel empowered and supported to do so.

An important aspect of this culture is ensuring equality of status and seniority between trading teams (the risk-takers) and those tasked with policing them. Disparities in status can be overt, such as differences in job title, grade, tenure or salary. They can also be more subtle: for example, the company may promote a culture that glorifies risk-takers and devalues compliance professionals or normalises exceptionalism for individuals whose roles involve pushing boundaries.

Evidence from T&S professionals suggests that more needs to be done to create a Culture of Challenge within tech companies. This is particularly important given the high status afforded to teams responsible for product development or recommender systems and the perception of safety as a drag on innovation.

 **Ensure T&S professionals have the authority to proactively surface risk and to implement safety initiatives across all aspects of the service and product design and delivery through professional standards and regulation.**

 **Require companies to keep written records of advice from T&S to senior leadership, including when leadership ignores warnings or rejects recommendations.**

## 2.3 Reporting lines

—

“To whom Trust & Safety reports internally can also directly or indirectly affect its ability to have a voice independent of competing incentives, such as revenue, public policy, or public relations considerations...”

Reporting directly to the CEO conventionally signals the team’s importance and that Trust & Safety is a major priority for the company.

A Trust & Safety leader who reports to the COO can indicate that Trust & Safety is tightly aligned with the company’s business needs and objectives but may also suggest that if Trust & Safety needs conflict with revenue generating interests, Trust & Safety needs may be deprioritized.

Reporting up to the CLO can mean the Trust & Safety team is more compliance-focused and may not directly influence new and emerging products and services.”

INTRODUCTION TO TRUST & SAFETY, TRUST & SAFETY PROFESSIONAL ASSOCIATION (TSPA)<sup>92</sup>

—

Concretely, this suggests that the person leading T&S should report directly to the CEO or Board.

Alternative reporting lines may also constitute a conflict of interest. For example, whilst T&S may support legal compliance efforts, if T&S leaders report to a Chief Legal Officer who is also responsible for minimising a company’s legal risk (for example, civil claims for harms experienced by users or regulatory investigations into the safety of a product or service) it is foreseeable that T&S’s need to proactively surface, scrutinise and discuss risk of harm may conflict with legal priorities, by increasing legal exposure.

More widely, research commissioned by the UK Government found that T&S leaders are concerned that Trust & Safety is perceived as symptomatic of regulatory changes and challenges, rather than an essential part of a business’ mission to protect its customers.<sup>93</sup> Currently, coverage of legally enforceable minimum safety standards

across the globe is inconsistent. Therefore, a safety strategy that is exclusively determined by compliance requirements would not comprehensively address risk of harm globally.

Reporting to external affairs leads presents another potential conflict. Evidence from T&S professionals show that these teams are frequently valuable allies when lobbying leadership to enhance safety. However, team priorities may diverge if T&S's duty to surface risk of harm creates the spectre of reputational risk for the company or undermines efforts to resist or modify legislative proposals that would fetter the company's commercial freedom or jeopardise its business model.

---

**“One of the things that has been raised is the fact that, at Twitter, the team responsible for writing policy on what is harmful reports separately to the CEO from the team that is responsible for external relations with governmental officials. At Facebook, those two teams report to the same person. The person who is responsible for keeping politicians happy is the same person who gets to define what is harmful or not harmful content.”**

FRANCES HAUGEN, FACEBOOK WHISTLEBLOWER<sup>94</sup>

---

In addition to these tensions, there is some evidence of reporting structures that simply fail. For example, Meta's global Trust & Safety lead could not articulate who was responsible for responding to research findings on harm and wellbeing.<sup>95</sup> Similarly, both Frances Haugen and Arturo Béjar reported that their concerns hit a glass ceiling and were effectively returned to sender with no action taken. Again, as the Wall Street Journal reported, Meta launched its privacy shield ignoring warnings from its safety experts that it would provide cover for paedophiles and was a “recipe for disaster”.<sup>96</sup>

Given the importance of T&S's work, the most appropriate reporting structure is for T&S leadership to report directly to the CEO. This arrangement makes it harder for CEOs to ignore, hide from or minimise warnings about risk. It is worth noting that a direct reporting line into the CEO is not a panacea. It must be supported by:

- a written mandate to act in the best interests of safety – even when these conflict with the business interests of the company;
- the right to veto or delay proposals which create safety risks that cannot be

effectively mitigated; and

- governance processes to document T&S advice and leadership’s response, including instances when advice is not taken.



**Establish a direct reporting line from T&S leadership to the company’s CEO/ senior leader.**

## 2.4 Safety Metrics



**“For Meta, a problem that is not measured is a problem that doesn’t exist.”**

ARTURO BÉJAR, INSTAGRAM WHISTLEBLOWER<sup>97</sup>



Tech companies are data driven, and progress is measured by tracking key data points (metrics). If a company does not have metrics relating to safety or a specific safety issue, then it is highly unlikely that the issue will be prioritised. Even when efforts are made to address a safety concern, without accurate metrics, it is impossible to know whether those efforts are working.

Furthermore, if a company chooses to measure and track poor safety metrics, targets may be achieved with little or no impact on user safety. Reporting incomplete metrics and not reporting relevant metrics can create a misleading impression about the safety of a service or product. This makes it hard for policy makers, regulators, and parents to form an accurate view on risk.

For example, evidence from whistleblower, Arturo Béjar, indicates that in 2021, data was either not being routinely collected or, when it was being collected, was not being acted upon or reflected in the company’s transparency reports.<sup>98</sup>



**“I witnessed firsthand how the different teams in Wellbeing were unable to do research or deploy products because of a corporate culture that does not want to understand the harm its products enable.”**

ARTURO BÉJAR, INSTAGRAM WHISTLEBLOWER<sup>99</sup>



In 2024, Arturo Béjar reported to the European Commission that “*the lessons Meta has taken from whistleblowers and information leaks has been to change internal practices to prevent reputational damage rather than to take responsible actions toward mitigation. Internal information has been locked down*”.<sup>100</sup> One such lesson was to stop minuting some meetings.<sup>101</sup> Additionally, a shareholder initiative to require Meta to publish data on harms to children in its annual report was blocked in line with the recommendation from the company’s Board.<sup>102</sup>

Beyond Meta, a review of the latest transparency reports from YouTube,<sup>103</sup> TikTok,<sup>104</sup> X,<sup>105</sup> and Snap<sup>106</sup> shows they all focus on content removal and do not (for example) include metrics on recommender systems, time spent on platform, take up of safety tools or the efficacy of their strategy to enforce their minimum age policies.<sup>107</sup>

**When T&S teams are limited in what data they are allowed to collect, how they measure progress, what metrics are prioritised by the company and how companies report data, their ability to surface, understand, and manage risk is severely constrained.** If, as Arturo Béjar suggests, heightened scrutiny leads to a defensive culture where access to data is locked down, these challenges are further exacerbated.

T&S teams must be free to determine what data it collects and shares to address safety issues. Data and metrics must also be shared transparently so that external stakeholders have an accurate picture about how safe — or unsafe — a product or service is.

 **Include meaningful and consistent safety metrics in company-wide targets with the same status as other business goals.**

## 2.5 Safety by design

---

“Companies, press, and regulators focus on content moderation as a solution to harmful speech, but this obscures the more structural cause: platform design and business models that encourage and invite harm.”

GRADY BERRY, AUTHOR OF THE FOCUS ON FEATURES PROJECT AND FORMER SENIOR SOFTWARE ENGINEER AT GOOGLE<sup>108</sup>

---

Safety by design seeks to address known or anticipated risk of harm upstream through product design. The aim is to prevent or substantially reduce the risk of harm occurring in the first place. Age-appropriate design reflects the rights of children under the UN Convention on the Rights of the Child and anticipates the vulnerabilities, capacities, and needs of children at different development stages.<sup>109</sup>

Content moderation is an important aspect of T&S work. Tech companies need effective systems to respond to harm once it arises including detecting and removing harmful content. However, managing harm retrospectively is not sufficient. Products and services must be safe and age appropriate by design and by default.

—

**“When Trust & Safety is going well, no one thinks about it or talks about it. And when Trust & Safety is going poorly, it’s usually something that leadership wants to blame on policies. Quite frankly, policies are going to be a Band-Aid if your product isn’t being designed in a way that actually doesn’t encourage abuse.”**

DEL HARVEY, FORMER HEAD OF TRUST & SAFETY, TWITTER<sup>110</sup>

—

Both safety by design and age-appropriate design require service providers to consider safety throughout the lifecycle of a digital service or product –during development, deployment, and retirement. The extent to which T&S professionals are embedded into product and service development varies from company to company.<sup>111</sup> However, TSPA’s *Introduction to Trust & Safety*<sup>112</sup> suggests that integrating safety by design and age-appropriate design into product development is not yet a core aspect of many T&S teams’ work.

**Research commissioned by the UK government found that T&S teams are consulted by products teams at the design phase only 52% of the time.** More commonly, they are likely to be consulted at the pre-deployment phase, when a product is built and ready to launch. Those T&S professionals consulted at this stage said that pre-deployment was too late.<sup>113</sup>

When asked what aspects of the product lifecycle (development, governance, enforcement, improvement or transparency) they found most challenging, 50% of T&S professionals working at large platforms identified product development as the most difficult. The other 50% cited governance. This suggests that T&S professionals at major online services face systemic barriers rather than technical or operational ones when advancing safety priorities.<sup>114</sup>

---

“...Overall, over time, over multiple instances, I and other people at the company felt that there was a pattern of putting more pressure to ship and compromising processes related to safety and problems that happened in the world that were preventable.”<sup>115</sup>

WILLIAM SAUNDERS, FORMER OPENAI EMPLOYEE AND SIGNATORY TO THE RIGHT TO WARN<sup>116</sup>  
OPEN LETTER

---

Achieving full integration of T&S into product and service design would involve several steps, such as:

- routinely including T&S professionals in product development teams;
- giving T&S the power to veto, adapt, postpone, or withdraw products, services, and features where safety issues have not been sufficiently mitigated;
- providing training on safety issues for non-safety team members such as designers and engineers so that they can spot issues and seek advice;
- creating product roadmaps that factor in the need to carry out safety tests and reviews – even if this disrupts launch plans.

---

“The GDPR has had a dramatic effect on how tech companies approach privacy. Product teams have to think about privacy at every stage of product development, they receive training on privacy by design principles and privacy specialists are included in sign off procedures as routine. A similar thing needs to happen with safety - rather than fixing things when they go wrong and building safety tools in silo, T&S needs to be fully integrated into all aspects of product strategy from the outset.”

VAISHNAVI J, FOUNDER OF VYANAMS STRATEGIES AND EX-META, GOOGLE AND TWITTER<sup>117</sup>

---

 **Embed T&S into all stages of design and development processes, granting T&S professionals the authority to modify, delay, halt, or withdraw services, products, algorithms, features, and functionalities if safety risks have not been sufficiently mitigated.**



## 2.6 Resourcing and capacity

—

**“Ask anyone in Trust & Safety, it’s like, we need money and people to do this.”**

ERIC HAN, EX GOOGLE, TWITTER, UBER, AND TIKTOK<sup>118</sup>

—

For T&S to operate effectively, it must be fully resourced. This includes not only sufficient budget and human resources within T&S but also access to other teams, such as engineering, product, and marketing, with whom collaboration is essential to build and implement safety improvements. Whilst tech companies have a reputation for intense work cultures, accounts from T&S professionals suggest safety teams are often overstretched.

—

**“In an environment where the scope and scale of abuse and the subsequent work it creates is so vast, it’s easy to take on more than you are scoped or resourced to do. This may stem from leadership encouraging you to take on more, or from your team’s desire to solve problems they identify - even if it’s not necessarily their job to fix them. However, you simply will not be able to do it all, and trying to accommodate everyone and everything will result in your team overloaded, overwhelmed, and potentially derailed from their mission or goals.”**

LEADERSHIP ADVICE FOR NEW TRUST & SAFETY LEADERS, INTEGRITY INSTITUTE<sup>119</sup>

—

An investigation by Bloomberg into Roblox included reports from a T&S professional who stated that her team could not keep up with hundreds of escalated child safety reports – far more than they could realistically address. In addition the news outlet reported: *“Eight current and former Trust & Safety workers say user growth at Roblox takes priority over child safety. They describe calls for more resources going unanswered, resulting in a backlog of incident reports and the departure of one manager who left after promises for extra staff went unfulfilled”*.<sup>120</sup>

Since 2022, substantial layoffs of T&S professionals have been reported. In some cases, teams and roles have been dismantled altogether. In its report, *Big Tech Slideback: How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections*, non-profit media watchdog Free Press concluded:

*“In 2023, the largest social-media companies have deprioritized content moderation*

*and other user Trust & Safety protections... These companies have also laid off critical staff and teams tasked with maintaining platform integrity... this has created a toxic online environment that is vulnerable to exploitation.”<sup>121</sup>*

In November 2024, the Australian e-Safety Commissioner reported that since Elon Musk acquired Twitter (now X) in 2022, he had laid off one third<sup>122</sup> of the platform’s safety team, including 80% of its engineers.<sup>123</sup> EU Justice Commissioner, Didier Reynders, described Twitter’s layoffs as a “source of concern”.<sup>124</sup>

---

**“I find that [the layoffs] so baffling that I wouldn’t even know where to start. I mean, those folks were just good people doing their best to keep people safe. And it’s unclear why you would target that work for removal, especially more recent [layoff] rounds when the folks targeted were ones who work on misinformation and election integrity.”**

DEL HARVEY, FORMER HEAD OF TRUST & SAFETY, TWITTER<sup>125</sup>

---

Mark Zuckerberg declared 2023 the “year of efficiency” and initiated widespread layoffs across Meta. Since then, tech companies, including Alphabet, Amazon, Microsoft, X, Discord, Snap, and TikTok have made cuts to T&S.<sup>126</sup>

---

**“I don’t think we should reflexively think that having fewer Trust & Safety workers means platforms will necessarily be worse... However, many of the people I’ve seen laid off are amongst the most thoughtful in rethinking the fundamental designs of these platforms, and if platforms are not going to invest in reconsidering design choices that have been proven to be harmful — then yes, we should all be worried.”**

RAVI IYER, FORMER META PROJECT MANAGER<sup>127</sup>

---

While protecting share price and increasing profitability are key duties of corporate executives, the significant cuts to T&S teams may have come at a cost to user safety.<sup>128</sup>

Decisions about T&S should be determined by the level of risk a service poses, not market forces.

➤ **Allocate appropriate resources to safety teams, based on T&S leadership’s assessment of needs and safe operating capacity.**

➤ **Establish enhanced procedures when tech companies make staffing cuts that may impact safety, including mandatory risk assessments and a requirement to notify relevant regulatory authorities.**

## 2.7 Working environment

—  
“...the content itself, having to deal with that, having the mindset of having to think through these issues a lot, and constantly having to worry about what’s going on with society... it’s a lot of emotional pressure to deal with.”

TRUST & SAFETY IN EVERYDAY TECHNOLOGIES<sup>129</sup>

—  
Work culture plays a critical role in the ability of T&S professionals to keeping users safe. The mission-driven mindset of T&S professionals, the barriers they encounter when pushing for change, and the high-intensity nature of their working environment place significant pressure on T&S team members and leaders. This not only makes it harder for them to be effective, but, in some cases, also leads to burnout.

### Exposure to harm and trauma

Irrespective of their role, exposure to harmful and traumatising content is routine for most T&S professionals. This content includes child sexual abuse material, extreme violence (including sexual violence), and hateful and discriminatory speech and actions. Without adequate organisational safety systems and processes to support them, T&S teams describe being put in the position of first responders with inadequate backup or support.<sup>130</sup>

Failure to prioritise harm reduction can be especially difficult to accept for those who see horrendous material on a daily basis such as content moderators. Some

moderators have sued tech companies for trauma caused by unsafe working conditions. Companies who have been subject to claims from moderators include Meta,<sup>131</sup> Google,<sup>132</sup> TikTok,<sup>133</sup> OpenAI,<sup>134</sup> and Reddit.<sup>135</sup> Some claims are still ongoing, others have resulted in substantial settlement payouts.

Perhaps in light of such claims, companies have increased their focus on their duty of care to protect the health and safety of T&S professionals. Implemented measures include offering employee wellness programmes and access to help lines. However, addressing the structural barriers that T&S professionals face daily to safeguard users, citizens, and society might more effectively support their mental health.

As noted earlier, content moderators working in the Global South are disproportionately at risk. In 2023, research published by academics at Middlesex University on the psychological impact of content moderation on content moderators found that:

*“despite their centrality to the business model of many online platforms (Barrett, 2020) moderators remain an undervalued, often hidden profession (Gillespie, 2018), frequently located in the global south and commonly not afforded the same benefits and provisions as offered to other professionals (Roberts, 2019; Jereza, 2021). However, the increasing amount of literature demonstrating the potential negative consequences of this work (e.g., Benjelloun & Otheman, 2020; Roberts, 2019) illustrates that companies need to take more responsibility for the welfare of these employees and volunteers and provide better protection against the possible deleterious psychological effects.”*

Interviews with T&S professionals conducted by Shulruff reveal that *“professionals experience stress not only from exposure to traumatic content and bad behaviour, but also from the pressures of standing between harms and the people they are working to protect”*.<sup>136</sup>

In addition to providing a safe working environment, tech companies could make the choice to protect their employees and subcontractors, by taking an upstream approach to risk reduction. **Prioritising harm prevention through responsible design, rather than focusing safety efforts on detection to safeguard business models, has a positive impact on service users and also on the health and safety of frontline T&S professionals.**

## T&S: risk of exposure to psychological harm

BY DR RICHARD GRAHAM, CONSULTANT PSYCHIATRIST AND GLOBAL EXPERT  
IN ONLINE HEALTH, SAFETY AND WELLBEING

In 1971, following the rapid growth in commercial air travel, US air traffic controllers were commonly reporting “vocational ‘burn out’” Many air traffic controllers had served in the military and were used to working under extremely challenging conditions.<sup>137</sup> However, they were ultimately overwhelmed by the increasing complexity and volume of work and the failure (or lack) of systems and processes needed to carry out their duties effectively and safely.

A series of fatal mid-air collisions led to change in working practices and prompted research into burnout, and its impact on both physical and mental health. There are many obvious parallels to the stresses that Trust & Safety Professionals face today, including that the consequences of impaired decision making could have global impacts. Yet compared to air-traffic controllers in the 1970s, the systems and processes within which Trust & Safety Professionals work and the consequences for their own and other’s safety are under considered.

In addition to occupational stress, both acute and long term, like many first responders, Trust & Safety Professionals (including professionals such as data labellers for the training of AI), will be exposed to distressing and illegal content (e.g. images of child abuse) and conduct (e.g. grooming). The extent of exposure will vary, but in certain situations it could lead to Post-Traumatic Stress Disorder (PTSD). A recent study on the mental health, secondary trauma experiences and wellbeing of content moderators concluded:

*“There was a dose-response effect between frequency of exposure to distressing content and psychological distress and secondary trauma, but not wellbeing.”<sup>138</sup>*

As with other first responders, such as firefighters or paramedics, working conditions can play a significant role in supporting them, and in reducing the impact of stress and exposure to potentially traumatic content. In the same research, the authors conclude:

*“The results suggested supportive colleagues and feedback about the importance of their role ameliorated this relationship.”*

In the absence of fairness, proper prioritisation, support, recognition, or rewards (and not just financial ones), Trust & Safety professionals face additional risks, including the possibility of developing “moral injury”.

*“Moral injury is understood to be the strong cognitive and emotional response that can occur following events that violate a person’s moral or ethical code... [it] can cause profound feelings of shame and guilt, and alterations in cognitions and beliefs (e.g. ‘I am a failure’, ‘colleagues don’t care about me’), as well as maladaptive coping responses (e.g., substance misuse, social withdrawal, or self-destructive acts).”<sup>139</sup>*

Further,

*“(Moral injury) Can contribute to the development of mental health problems, including depression, PTSD and anxiety.”<sup>140</sup>*

The risk of moral injury is heightened in certain situations familiar to Trust & Safety Professionals, notably when there is loss of life of a vulnerable person (e.g. child, woman, elderly); if leaders are perceived to not take responsibility for the event(s) and are unsupportive of staff; if staff feel unaware or unprepared for emotional/psychological consequences of decisions; if there is a lack of social support following a potentially morally injurious event (PMIE).<sup>141</sup>

Given the complex structures and differing priorities within tech companies, it would appear that experiencing a PMIE may be highly likely for many working in Trust & Safety. It is worth repeating that social support can still play a key role in prevention. But there is a risk that some will develop a relatively new clinical condition, Post-Traumatic Embitterment Disorder:

*“The term ‘posttraumatic embitterment disorder’ (PTED) was recently introduced, (and is) characterized by prolonged embitterment, severe*

*additional psychopathological symptoms and great impairment in most areas of life in reaction to a severe negative but not life-threatening life event.”*

At this point we cannot know how many of those working in Trust & Safety may be at risk of developing PTED, but whether it is stress, burnout, PTSD or PTED, the implications are that Trust & Safety Professionals require working conditions that both support them and protect them from the potential harms that may arise in the course of their work. Whilst their health and wellbeing require consideration from an ethical and even employment law perspective, there is one other perspective to consider.

In the global race to regulate online services, much emphasis is placed on the moderation of content and user conduct. If the professionals charged with undertaking that complex work at scale are not in good health and functioning well, their ability to create safer online environments will be compromised – much like the air traffic controllers in the 1970s. Law firms will take note.

## **Burnout**

For T&S professionals, the combination of continued exposure to harmful content, insufficient structural support, and the scale and nature of risk being managed can be overwhelming. Some T&S professionals thrive under these conditions; some struggle. Others thrive for long periods of time before becoming overwhelmed.

---

**“Even with setting boundaries and taking care of yourself, it’s more likely than not that you and your team will experience burnout at some point.”**

LEADERSHIP ADVICE FOR NEW TRUST & SAFETY LEADERS, INTEGRITY INSTITUTE<sup>142</sup>

---

Research conducted with content moderators in the UK found that their daily targets for reviewing videos were unmanageable, often leading to mistakes which in turn

impacted their bonuses and career progression.<sup>143</sup> Moderators also reported being unable to take sufficient time off after viewing highly distressing content and that the therapeutic support was limited and difficult to access due to changing shift schedules.

## External perception

—

**“[T&S professionals] find themselves out of place within the corporate cultures which do not elevate or affect approaches that are more risk-conscious.”**

TOBY SHULRUFF, ARIZONA STATE UNIVERSITY<sup>144</sup>

—

The challenge that T&S professionals describe when their interests and priorities differ from those of their employers may be exacerbated by the external view of tech companies as homogenous entities. As a result, T&S professionals may find themselves under fire from both directions: they encounter internal resistance when making the case for enhanced safety standards, whilst also facing legitimate criticism from safety advocates when tech companies fail in their responsibility to keep users safe.<sup>145</sup> This dynamic may place T&S professionals in an uncomfortable position, particularly if they are expected to act as spokespeople or play a role in promoting or defending safety strategies that they believe to be inadequate or partial.



**Develop, and consistently apply, industry-wide minimum health and safety standards to protect the physical and mental health of all T&S professionals — irrespective of geography or contract type.**



**Enhance support and protections for those who raise safety concerns, including by codifying the four principles of the Right to Warn<sup>146</sup> across the entire tech sector.**



## 2.8 Conflict of interest

---

“Protecting users and ensuring trust in products may come into conflict with other company objectives, such as product growth and marketing, as well as company or societal values.”

INTRODUCTION TO TRUST & SAFETY, TRUST & SAFETY PROFESSIONAL ASSOCIATION<sup>147</sup>

---

A conflict of interest arises when an individual has competing interests that are, or may be, incompatible. Such conflicts can be real or perceived. Even when a conflict of interest has no practical impact on how an individual operates or the decisions they make, perceived conflict of interest undermines trust and credibility. It can also be a source of anxiety when individuals are expected to navigate and manage conflicts without a clear framework or rules to guide them.

Tech company employees are required to carry out their duties in the best interests of the company. This duty is overridden when an employee becomes a whistleblower, but most employees do not become whistleblowers.<sup>148</sup> As analysis of T&S job postings shows (see above), the primary requirement of a T&S professional is to make the service safer for users. Nonetheless, as employees they must also work in the best interest of the company, as set by the CEO and Board. This dual obligation creates the potential for a conflict of interest, which could be addressed by developing a professional code of practice that explicitly confirms that the primary duty of T&S professionals is to act in the best interests of society, citizens, and users.

### Remuneration

Like other employees at tech companies,<sup>149</sup> some T&S professionals may be given the chance to participate in share schemes.<sup>150</sup> The value of these schemes can be significant, and participation means the individual has a vested interest in the company's profitability. While T&S professionals should not be less well paid than their colleagues in other departments, the potential for conflict of interest could be resolved by creating alternative schemes of equivalent value that are connected to safety rather than profitability or growth targets and/or by including conflict management within a professional framework.

 **Provide differentiated — but equivalent — remuneration and incentivisation schemes to protect the independence of T&S professionals.**

## 2.9 Nascency of the profession

—

**“We need to create space to hear from people on the front lines. We need to give them protection so they can share their experiences. Only then can we begin to understand the full scope of the problem and find solutions”**

ANIKA COLLIER NAVAROLI, FORMER SENIOR EXPERT, TWITTER’S US SAFETY POLICY TEAM<sup>151</sup>

—

### Professional bodies

Established professions typically create professional bodies to support and oversee the work of practitioners as they move through their careers. Roles that professional bodies can perform include drafting and publishing best practice advice and guidance,<sup>152</sup> providing training and accreditation schemes;<sup>153</sup> setting and enforcing professional standards;<sup>154</sup> thought leadership and policy development;<sup>155</sup> representing members’ interests to policy makers;<sup>156</sup> organising collective action and negotiating collective agreements,<sup>157</sup> and providing practical and mental health support to members.<sup>158</sup>

In recent years, T&S has begun to establish its professional identity. This development is reflected in several areas, for example:<sup>159</sup>

- the emergence of professional organisations and membership bodies such as the Trust & Safety Professional Association (TSPA), the Integrity Institute and All Tech Is Human;
- the proliferation of podcasts and online forums devoted to discussing the profession and sharing challenges, best practices, and career experiences;
- the growth of convening events, most notably TrustCon, held annually in San Francisco with satellite events in Europe and Asia;
- efforts to standardise training, ranging from peer-to-peer guidance to university courses.

The Integrity Institute is an independent think tank that enables T&S professionals to provide input into policy development and public discourse issues independently of the companies they work for. In the run up to the Senate Hearings on child safety with tech CEOs in January 2024, its members and fellows supported senators by preparing briefing decks for staffers, proposing questions for senators to ask, and writing blogs with policy recommendations.<sup>160</sup> The Integrity Institute has also engaged with policy makers and regulators in Europe on the Digital Services Act and in the UK in preparation for the introduction of the Online Safety Act.

Beyond the Integrity Institute, the focus of membership organisations appears to be on the sharing of best practice, identity formation as a new profession, and peer-to-peer learning and support. These efforts are all critical and the response<sup>161</sup> to events such as TrustCon shows how much T&S professionals value these resources and opportunities to convene, learn, share, and support each other.

The current scope of T&S membership organisations does not appear to extend to carrying out, or assessing the value of, the following:

- creating professional standards and oversight mechanisms (see Professional Standards below);
- defining the skills and qualification requirements for different roles and responsibilities and developing training programmes;
- facilitating collective action and negotiations on issues such as pay and working conditions;
- developing and implementing strategies to advocate for T&S's independence and authority to act in the best interests of users.

## **Professional standards & oversight**

Codes of practice (also referred to as codes of conduct, codes of ethics, and professional standards) are rules or principles that guide the conduct of professionals working in a specific industry or field. These codes are especially valuable for professions whose work has a public interest element, including public safety, or where professionals are required to navigate conflict of interest (for example accountants who are paid by companies to audit and sign off on their accounts). Professions that adhere to codes of practice include lawyers,<sup>162</sup> medics,<sup>163</sup> social workers,<sup>164</sup> law enforcement officers,<sup>165</sup> pilots,<sup>166</sup> psychotherapists,<sup>167</sup> accountants,<sup>168</sup> and engineers.<sup>169</sup>

Whilst it may seem counterintuitive, **introducing professional standards and oversight is in the best interests of those who are subject to them because it enables them to push back if they are asked – either implicitly or explicitly – to operate outside prescribed professional boundaries.** There is therefore a strong case to be made for T&S professionals to work together on developing codes of practice.

For a code to be robust and garner widespread support, it should be developed by T&S professionals in consultation with stakeholders such as academics, civil society, those with lived experience of harm, and end users, including children. Grounding these standards in, or making reference to, existing frameworks such as human and child rights conventions, ESG requirements and (where they exist) regulatory provisions on governance, systems, and processes for online services would further enhance their impact and adoption.

While codes of practice may be voluntary, they are most effective when enforced through self-regulatory, co-regulatory, or regulatory frameworks. Under such frameworks, failure to uphold standards has consequences such as disciplinary action, fines, suspension, or even revocation of the right to practice. For the wider profession, however, they enhance trust and confidence, provide a framework to navigate conflicts of interest, and empower practitioners to say no or to insist on change by formalising the duty to serve and protect.




## **Professional training and qualifications**

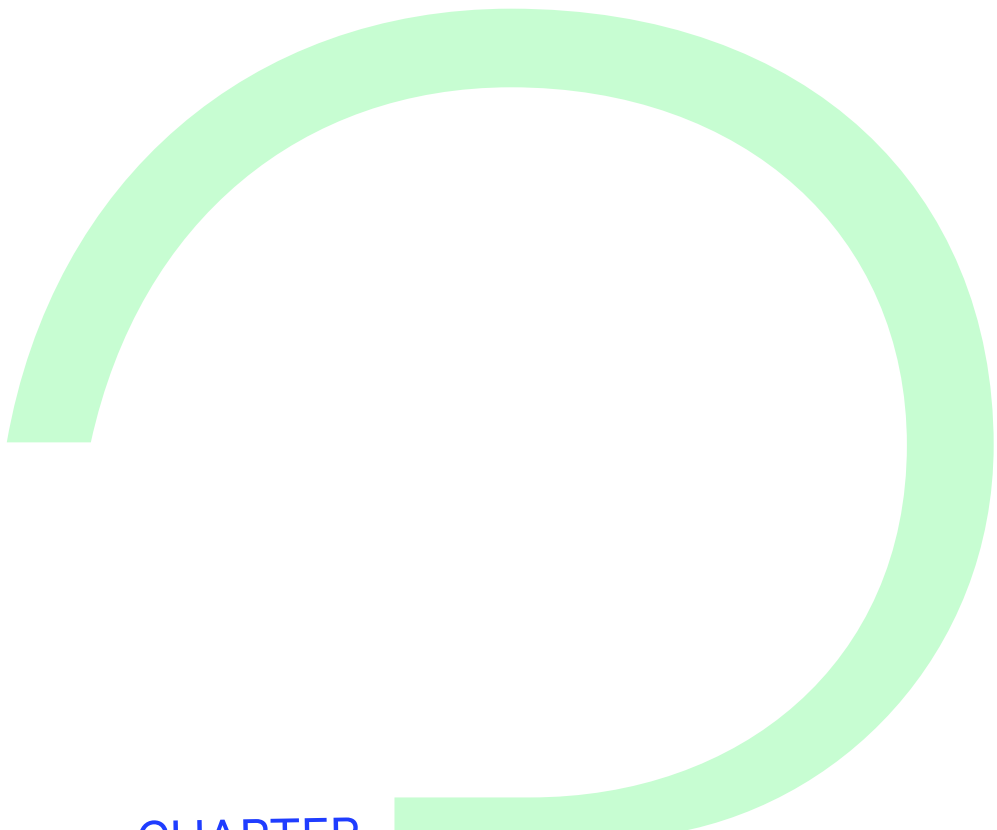
One of T&S's strengths is that professionals come from varied backgrounds, bringing with them invaluable subject or regional expertise that enhances their ability to do their job. However, this diversity also means that there is not yet clarity on the specific skills needed or consensus on the minimum qualifications required for particular roles. Given the importance of the decisions made by T&S, greater clarity on what the job is, who has the skills and expertise to do it, and how these skills are attained, tested, and monitored is needed.

For example, there is no consensus on how long content moderators should be supervised as trainees before making independent decisions on content removal. Similarly, there is a lack of consensus on the requirements for continued professional development (CPD) or the additional qualifications or training needed for professionals promoted to positions of greater responsibility.

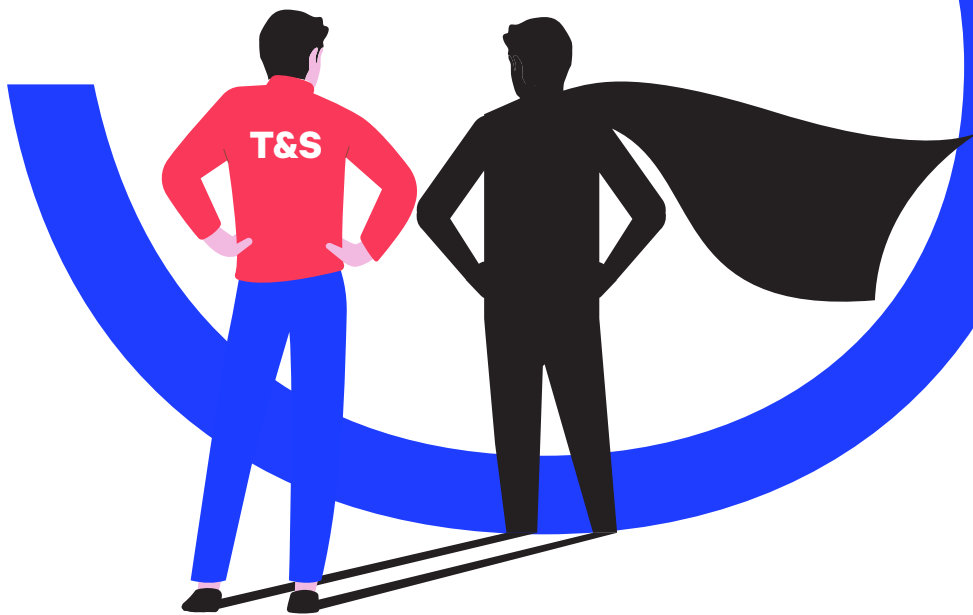
Whilst some courses have been introduced to prepare individuals for entry-level T&S roles, these are neither compulsory nor standardised. They are also predominantly available in the US.

The mandate and authority of T&S would be strengthened by defining the minimum skills and expertise requirements for entry to the profession and for specific roles; setting requirements for continued professional development; ensuring new practitioners are given necessary training and support before they operate independently; and creating systems and processes for continued monitoring and evaluation of critical competencies.

-  **Expand the remit of T&S membership organisations to include representing the needs, interests, and concerns of T&S professionals to policy makers, regulators, and industry.**
-  **Develop professional codes of practice to guide and protect those working across the T&S ecosystem.**
-  **Clarify the skills and qualifications required for different roles within T&S and create validated training programmes to support new appointments and continued professional development.**



CHAPTER



---

**Over the past few years,  
I'm definitely becoming  
more and more optimistic  
about the long arc of our  
profession**

JEFF ALLEN, INTEGRITY INSTITUTE, FEBRUARY 2025<sup>170</sup>

---

# 3. Who can make change?

In the Introduction, we set out a series of recommendations to address the challenges described in Section Two. In this section, we identify key stakeholders who have a role to play in bringing about change and outline specific actions they can take.

## 3.1 Policy makers

—

“Policymakers should focus on measures that empower integrity workers to do their jobs better and have more influence within the companies as the most effective lever for making platforms safer in the long term.”

CHILD ONLINE SAFETY, INTEGRITY INSTITUTE<sup>171</sup>

—

The growing number of legally prescribed compliance obligations shifts user safety from a “nice to have” to a “must have”.<sup>172</sup> This is the most reliable way to raise safety standards. Online safety laws have been passed across Europe<sup>173</sup> and in many other countries and regions around the world. In the US where online safety regulation has yet to pass at a federal level and remains inconsistent at a state-level, it is notable that T&S professional bodies are working with politicians to support their efforts to legislate.<sup>174</sup>

- **Understand how the remit, authority, and resourcing of T&S within organisations impacts safety.**
- **Include a legal requirement to proactively surface, interrogate, and understand the risks of harm from digital products and services in online safety laws.** This means rewarding tech companies who encourage T&S to take a proactive approach to surfacing and responding to risk, while imposing penalties for those who don’t.



- **Enshrine T&S’s independence and authority to act in the best interests of services, users, and society in legislation.** This means introducing laws that require tech companies to have policies and processes on (for example) risk review, reporting lines, whistleblowing and conflicts of interest.
- **Ask detailed questions.** You may find the questions in the T&S assessment tool at Chapter Five helpful.
- **Require tech companies to be more transparent about T&S** and to provide information about the systems, processes and governance practices that support T&S’s work.

## 3.2 Regulators

—

“There is the opportunity for regulation to empower Trust & Safety teams on the inside, to help enable them to be more effective.”

JEFF ALLEN, INTEGRITY INSTITUTE<sup>175</sup>

—

This report details industry-wide failures to establish and maintain the systems, processes and governance structures required to facilitate T&S’s role in safeguarding society, citizens, and users. Risk-based regulatory frameworks that require tech companies to create safe and age-appropriate services and products by design and by default must include measures on T&S.

- **Understand how the remit, authority, and resourcing of T&S within organisations impacts safety.** Like policy makers, regulators must understand the ways in which the systems, processes, and governance structures under which T&S operates can mitigate or amplify safety risks.
- **Consider how the recommendations in this report will inform your regulatory approach, including risk assessment and transparency requirements.** For example, in the UK Ofcom’s draft Illegal Harms Code includes a requirement for regulated services to have written statements of responsibilities for senior members of staff involved in making decisions related to the management of online safety

risks.<sup>176</sup> This ensures clarity on who holds ultimate decision-making authority within companies.

- **Ask detailed questions.** You may find the questions in the T&S assessment tool at Chapter Five helpful.
- **Establish enhanced mechanisms for whistleblowing.** For example, under the Digital Services Act, the European Commission has created a Whistleblower Tool to facilitate anonymous reports.<sup>177</sup>

### 3.3 Tech companies

—

“Just as de-investing from T&S is a statement, I think investing will also be a statement.”

ALICE HUNSBERGER, VP OF TRUST & SAFETY, PARTNER HERO<sup>178</sup>

—

The evidence from T&S professionals in this report indicates that many CEOs could do more to empower and support T&S.

- **Recognise and uphold the right and responsibility of T&S professionals to act in the best interests of society, citizens, and users** – even when this conflicts with the commercial interests of the company. Use the T&S assessment tool in Chapter Five to evaluate current practices.
- **Provide T&S with a clear mandate** to proactively surface risk, to implement safety strategies, and require safety changes across all aspects of your service and product design and delivery.
- **Ensure your governance systems and processes reflect T&S autonomy**, including documenting instances when advice or warnings from T&S was ignored or overruled.
- **Be more transparent about T&S.** Provide information about the systems, processes and governance practices that support T&S’s work including how their efficacy is measured and monitored. Share best practice examples.

- **Establish safe and sustainable working practices for all T&S professionals.**
- **Support T&S professionals working within, or on behalf of, your company to pursue the recommendations in this report.** Provide resources (time and money). Do not place restrictions on T&S professionals’ right and ability to discuss these issues internally or externally. Do not penalise those calling for higher standards.

## 3.4 T&S organisations

—

“Content moderators show up every day and try to do their best in impossible situations. But the modern-day public conversation should not be susceptible to the whims of any one company or individual.”

—

ANIKA COLLIER NAVAROLI, FORMER SENIOR EXPERT, TWITTER’S US SAFETY POLICY TEAM<sup>179</sup>

—

T&S membership organisations have a critical role to play in addressing the challenges and advancing the Key Recommendation of this report described in this report.

- **Facilitate discussion about the status and future of T&S.** Leverage your convening powers to foster discussion amongst T&S professionals and wider stakeholders.
- **Establish professional standards.** Develop consensus-based standards that codify the autonomy, rights, and responsibilities of T&S professionals. These standards could make clear that T&S professionals’ overriding duty to act in the best interests of society, citizens, and users – especially vulnerable users such as children.
- **Establish training programmes and accredited qualifications.** Continue to develop consensus-based skills and training programmes tailored to various T&S roles. These should support those entering the profession at any level and ensure the continued professional development of practitioners.
- **Expand your membership.** Make it easy and affordable to join your organisation. Proactively recruit T&S professionals from underrepresented groups and regions.
- **Champion the rights and needs of T&S professionals.** Harness the power of collective action to advocate for the enhanced status and independence of T&S

to policy makers, regulators, and industry. Support your members to secure safe working conditions, fair pay, and a clear mandate on their right and responsibility to prioritise safety. Consider whether your funding model curtails your freedom to act in the best interests of T&S professionals.

## 3.5 T&S professionals

---

“Within industry, workers can build internal coalitions, focusing on network formation within organizations to maintain a focus on responsible technology practices.”

DEB DONIG, SIEGEL RESEARCH FELLOW AT ALL TECH IS HUMAN<sup>180</sup>

---

Evidence from T&S professionals and their membership bodies indicates they recognise that the systems in which they operate are not fit for purpose. This raises important questions: why do they choose to stay? And are they right to do so?

Not all T&S professionals stay – some walk away and a brave few choose to speak out. Evidence on why T&S professionals stay is thin. Those working in T&S describe the work as interesting and a career in T&S can offer a good salary.<sup>181</sup> Beyond these practical considerations, T&S professionals also exhibit a “first-responder” mindset – that is to say, they describe feeling a sense of duty to stay and to protect despite the lack of agency, emotional toll,<sup>182</sup> and risk of burnout. As accounts from whistleblowers show, speaking out requires significant personal fortitude and carries legal, financial, professional, safety, privacy, and reputational risk.<sup>183</sup> Given these risks and their desire to protect, staying and playing a role in making online spaces safer may feel like the best option or the only manageable one.<sup>184</sup>

This report makes the case for minimum standards on T&S’s working conditions, its mandate to act in the best interests of society, citizens and users and the systems and processes that facilitate this work. That is to say, to enable T&S professionals to carry out their work safely and effectively within tech companies.

- **Understand your identity as a Trust & Safety professional.** Most T&S professionals focus on solving safety challenges rather than thinking about the status or purpose of their profession. The resources in the final section of this report offer valuable insights into understanding your identity as a T&S professional and how

it can shape the way you approach your role. The T&S assessment tool in Chapter Five will help you to evaluate your company's current approach against proposed best practice.

- **Get involved.** Find or start a group within your company, participate in T&S forums online, join a membership organisation, or unionise.
- **Advocate for change.** Consider campaigning for minimum standards and protections on pay and working conditions for T&S professionals worldwide. Push for T&S professionals' right and responsibility to act in the best interests of society, citizens, and users to be enshrined in law and formally recognised by tech companies. Support efforts to develop professional standards (e.g. codes of practice, training accreditations, conflict of interest guidance etc).
- **Leave.** This report makes the case for normative systems, processes, and governance to support T&S professionals. Responsibility for challenging poor practices should not fall disproportionately on individual workers. However, if you believe leadership's decisions are incompatible with the safety and best interests of society, citizens, and users, you may conclude that staying is untenable.

# 4. Trust & Safety resources

## Professional organisations and bodies

[Integrity Institute](#)

[All Tech is Human](#)

[Trust and Safety Professional Association](#)

[Trust and Safety Foundation](#)

[Digital Trust and Safety Partnership](#)

## Podcasts, newsletters and forums

[Trust in Tech](#)

[Journal of Online Trust & Safety](#)

[Safety is Sexy](#)

[Trust and Safety Mavericks](#)

[Impossible Tradeoffs with Katie Harbath](#)

[Quire](#)

## T&S conferences and webinars

[TSPA TrustCon, EMEA Summit and APAC Summit](#)

[Trust and Safety Forum](#)

[All Tech is Human](#)

[Trust and Safety Hackathon](#)

## Training and courses

[Trust and Safety Teaching Consortium, Stanford University](#)

## T&S Mentor Programmes

[All Tech is Human's Mentor program](#)

[TSPA Coffee Chats](#)

[T&S Mentor Match](#)

## Books & Articles

Sarah T. Roberts, (2019), [Behind the Screen: Content Moderation in the Shadows of](#)

Social Media, Yale University Press

Annalee Newitz, (22 November 2023), Trust and safety – the most important tech job you’ve never heard of, New Scientist

Toby Shulruff, (2024), Trust and Safety work: internal governance of technology risks and harms

## **T&S Glossaries**

Terminology

Key Roles & Functions

## **Safety by Design**

Australia’s e-Safety Commissioner, Safety by Design

5Rights Foundation and Digital Futures Commission (now Digital Futures for Children centre), Children’s Rights by Design

Thorn and All Tech is Human, Safety by Design for Generative AI: Preventing Child Sexual Abuse

Center for Human Technology

## **Whistleblower Information & Support**

Whistleblower Aid

Arturo Béjar, Instagram Whistleblower

Frances Haugen, Facebook Whistleblower

# 5. T&S Assessment Tool

This tool has been designed to enable stakeholders to assess current practices within a tech company against the Key Recommendations in Chapter One of this report. Tech companies can use it to understand where the policies, processes and systems meet proposed best standards and where they fall short. It can also be used to re-evaluate practices if changes are made in the future. T&S professionals may find it helpful as a prompt to evaluate the extent to which their right and responsibility to safeguard society, citizens and users are supported within the organisation they work for. Policy makers and regulators can use the assessment tool to encourage greater transparency and openness from tech companies.

Mandate and authority to act	NO	Partially	YES
We recognise that T&S professionals’ overriding duty is to ensure the safety of society, citizens and users — even when this conflicts with commercial priorities and interests. Our systems, processes and written policies reflect this.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Our T&S team has authority to proactively surface and measure risk of harm across all aspects of our service.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Our T&S team’s remit includes oversight of product design and AI systems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Our T&S team has the authority to modify, delay, halt, or withdraw services, products, features, and functionalities and AI systems if it determines that safety risks have not been sufficiently mitigated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Our T&S team has the freedom to collect, share, and record safety-related data and metrics it deems necessary to carry out its role.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Safety targets and metrics are included in business-wide goals.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



<i>Mandate and authority to act</i>	NO	Partially	YES
We are transparent about the systems, processes and governance practices we have put in place to support T&S's work. We share data about the efficacy of these measures in our Transparency report.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We require other teams to prioritise requests from T&S that relate to safety over requests from non-safety teams.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

<b>Governance systems and processes</b>	<b>NO</b>	<b>Partially</b>	<b>YES</b>
We do not make decisions that impact safety unless T&S has provided advice on risk to our CEO (or equivalent).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If advice from T&S is not followed, or only partially followed, by our CEO (or equivalent), we document the advice from all participants in the discussion and the reasons why T&S's advice was rejected.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When T&S team members engage with the work of legal and compliance teams, those teams provide clear advice upfront about their legal rights and duties.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dissenting views from T&S on legal and compliance assessments are recorded and disclosed to regulators.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T&S team members are not required support the company's defence of legal and regulatory proceedings.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We offer T&S professionals differentiated but equivalent remuneration and incentivisation schemes designed to protect their independence.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We recognise and support the four principles of the Right to Warn <sup>185</sup> and provide support and protections to anyone raising safety concerns.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Our procedures and policies apply to all T&S professionals working on our products and services including those working via sub-contractors and third-party vendors.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Governance systems and processes	NO	Partially	YES
Managers are trained to recognise, record, and respond to safety concerns raised by T&S professionals.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We take a proactive approach to identifying and responding to concerns raised by T&S professionals. We do not require T&S professionals to log formal complaints as a pre-condition of taking action.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We have procedures in place to protect the anonymity of T&S professionals raising complaints.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We actively promote a culture of challenge. We identify and address explicit and implied inequality of power between T&S professionals and the teams they collaborate with (e.g. legal, public relations and public affairs) and those whose work they oversee (e.g. product).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
We don't require T&S professionals to act as spokespeople or to sign off on communications or marketing materials relating to safety that they consider to be misleading or incomplete.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
To safeguard their independence, we have a written policy on the rights and responsibilities of T&S professionals.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The T&S policy includes managing real or perceived conflict of interests; advising leadership, legal and compliance teams; representing the company externally; raising complaints and remuneration including participation in share schemes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T&S policies are accessible to all T&S professionals, and we do not monitor or keep records of who has accessed the policies.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Safe and sustainable working practices	NO	Partially	YES
We allocate resources (financial and human) to safety teams based on the assessment of needs and safe operating capacity by T&S leadership.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Safe and sustainable working practices

NO

Partially

YES

Requests for additional resources are considered by our CEO (or equivalent). Decisions to refuse such requests are documented.



Decisions to include T&S teams in redundancy rounds are made by our CEO (or equivalent) and reported to the Board.



We have enhanced redundancy procedures when making staffing cuts that could impact user safety. These include conducting risk assessments, and a requirement to notify relevant regulatory authorities.



We have comprehensive health and safety policies, systems, and processes in place to protect the mental<sup>186</sup> and physical health of T&S professionals.



We recognise our responsibility to safeguard all those working in our T&S teams, irrespective of whether they are employed directly by us or via a third-party contractor.



We have full visibility on the health and safety policies, systems, and processes of all our T&S third-party contractors. We don't work with third parties who fail to meet our standards.



We regularly conduct anonymous surveys to understand whether T&S professionals feel supported, empowered, and able to raise concerns. Results are analysed by market, roles, seniority, and contract type.



Our executive leadership reviews the surveys and is responsible for agreeing on and implementing further action (if required).



We ensure transparency in job postings and interviews about the specific requirements of the role, so that candidates can make an informed choice about whether they are suited to working in T&S or ask for additional support they may need to perform the role.



## Professional standards and training

NO

Partially

YES

We have policies in place that describe the skills and qualifications needed to perform different roles within T&S and ensure that those appointed meet these requirements. If an employee doesn't meet such requirements, we provide additional training and support.



We provide support and training to lateral hires from outside the profession and to internal candidates who transfer into T&S. We also support T&S professionals when they take on additional responsibilities.



We have strategies in place to support continued professional development of T&S professionals at all stages of their career.



We support and encourage T&S professionals to convene and discuss the future of their profession.



We provide resources, time, and money to facilitate these discussions.



We do not place restrictions on the ability of T&S professionals to discuss their profession (including challenges) publicly.



We support the efforts of T&S professionals to develop professional standards, oversight and qualifications.



We do not penalise T&S professionals seeking to unionise or pursuing other forms of collective action.



We recognise the right of T&S professionals to withdraw labour.



# Endnotes

1. Integrity Institute, (04 April 2024), [We're at a tipping point as a profession.](#)
2. Open letter, (signed 4 June 2024), [A Right to Warn about Advanced Artificial Intelligence.](#)

## Chapter 1

3. [View of how to build a trust and safety team in a year](#), Volume 1 N°2 December 2022, Journal of Online Trust & Safety.
4. Integrity Institute, [What is an Integrity Worker?](#)
5. Trust & Safety Professional Association, [About Us.](#)
6. John Perry Barlow (February 1996), [A Declaration of the Independence of Cyberspace.](#)
7. Dibbell, J. (December 1993), [A Rape in Cyberspace.](#) Founded in 1990, LambdaMOO is a text-based virtual world, where users could interact in a self-governed environment. In 1993, Julian Dibbell documented a virtual sexual assault incident involving a character, Mr. Bungle, using a “voodoo doll” programme to attack two female avatars. This event, known as the “Bungle Affair,” highlighted early challenges in moderating virtual spaces. LambdaMOO’s new governance model required community consensus for disciplinary actions, limiting the administrators’ powers. During a meeting to address the incident, no consensus was reached, but an administrator eventually banned Mr. Bungle. However, he returned under a new persona. This incident prompted LambdaMOO to reconsider its governance structure, restoring administrators’ moderation powers three years later.
8. Stratton Oakmont, Inc. v. Prodigy Services Co., in [Dawn L. Hassell and Hassell Law Group, P.C., v. Yelp Inc. On Petition for Writ of Certiorari to the California Supreme Court](#),(p. 7).
9. See Part V of the Telecommunications Act 1996.
10. See Part V of the Telecommunications Act 1996 and also [The FCC’s Authority to Interpret Section 230 of the Communications Act \(October 2020\)](#), Federal Communications Commission.
11. Kosseff, J., (2019), [The Twenty-Six words that created the Internet](#), Cornell University Press.
12. Databite 134: Origins of Trust and Safety, (recorded on 07 August 2020), [Databite 134: Origins of Trust and Safety With Alexander Macgillivray and Nicole Wong](#), [Transcript], hosted by Robyn Caplan.
13. Facebook launched in 2004, YouTube in 2005 and, Twitter in 2006.
14. Klonick, K., (20 March 2017), [The New Governors: The People, Rules, and Processes Governing Online Speech.](#)
15. Databite 134: Origins of Trust and Safety, (recorded on 07 August 2020), [Databite 134: Origins of Trust and Safety With Alexander Macgillivray and Nicole Wong](#), [Transcript], hosted by Robyn Caplan. *“I think the other areas where things grew a lot were like having to deal with child sex abuse materials, was also one where the law was super clear, where nobody wanted to have that on their site anyway. So, you saw the development of teams both to identify and quickly remove, and the tooling that was necessary to do that generate very quickly in the Trust & Safety areas.”*
16. Rob Chesnut, (2022), [LinkedIn post.](#)
17. Ibid.
18. For example, the UK’s [Cyberbullying Taskforce](#) launched in 2016.
19. For example, Google’s [‘Be Internet Awesome’](#) digital literacy programme launched in 2017.
20. For example, YouTube launched parental controls in 2010. Similar tools were not launched on [TikTok until 2020](#) and [Snap and Instagram until 2022.](#)
21. For example, the [Christchurch Call](#) was signed in 2019.
22. For example, the ICT’s [Principles for the Safer Use of Connected Devices and Online Services by Children and Young People in the EU](#), was published in 2017.
23. [The Facebook files: A Wall Street Journal investigation](#), (2021), Wall Street Journal.
24. Trust & Safety Professional Association, (7 November 2022), [From academia to T&S | Lindsay, Aaron, and Kat](#), [Video], YouTube
25. Jahangir, R., (19 March 2024), [Amid Flurry of Online Safety Laws, the Global Online Safety Regulators](#)

Network is Growing, Tech Policy Press.

26. View of how to build a trust and safety team in a year, Volume 1 No.2 December 2022, Journal of Online Trust and Safety.
27. Spence, R., Bifulco, A., Bradbury, P., & Martellozzo, E., (2023), The psychological impacts of content moderation on content moderators: A qualitative study, Cyberpsychology Journal of Psychosocial Research on Cyberspace.
28. For example, Inside Facebook's African Sweatshop | TIME which reported that sub-contracted content moderators working on behalf of Facebook took home as little as \$1.50 per hour. Behind TikTok's boom: A legion of traumatised, \$10-a-day... | TBIJ which in 2022 reported that moderators employed to review content on TikTok by a contractor in Columbia work six days a week on day and night shifts, with some paid as little as 1.2 million pesos (£235) a month, compared to about £2,000 a month for content moderators based in the UK.
29. For those seeking to understand the experience and challenges of online content moderators, Radio 4's series 'The Moderators' is an excellent resource.
30. Shulruff, T. - Arizona State University, (March 2024), Trust & Safety in everyday technologies. See also Lehane, C., PhD, (4 June 2022), A Career in Trust & Safety: You know more than you know, Medium.
31. TSPA is a "*non-partisan membership association that supports the global community of professionals who develop and enforce principles, policies, and practices that define acceptable behavior and content online and/or facilitated by digital technologies. TSPA works to create and foster a global community of Trust & Safety professionals, collaborating with them to build a community of practice, and providing support as they do the challenging work of keeping online platforms safe.*" See Trust & Safety Professional Association, What we do.
32. Trust & Safety Professional Association, Key elements of a Trust & Safety team.
33. Trust & Safety Professional Association, Approaches to Trust & Safety.
34. UK Government's Department for Science Innovation and Technology, (October 2024), Technology and Trust and Safety: the state of play and integration of technology.
35. Trust and Safety Professional Association (TSPA), The Purpose and Role of T&S Teams.
36. Informed by The Business of Keeping the Internet Safe, JooHo Yeo (February 20024), LinkedIn.
37. Goggin, B., (29 March 2024), Social media companies pushed to improve child safety, NBC News.
38. Amazon, Amazon Brand Protection Report 2024.
39. Hickey, M., (26 March 2024), TikTok head of safety explains how app is pushing back against dangerous content, CBS News.
40. Everest Global, Inc., (2024), Everest Group Trust and Safety Services PEAK Matrix® Assessment 2024.
41. Between 2021-2023, safety tech firms raise US\$4.8 billion in investment. Source: Public - International State of Safety Tech Report: 2023 - Page 1 - Created with Publitas.com. See also Duco, Trust & Safety Market Research Report.
42. Trust & Safety Professional Association, (7 November 2022), From academia to T&S | Lindsay, Aaron, and Kat, [Video], YouTube.
43. Trust and Safety Professional Association, (10 May 2023), From Government to T&S | Brian, Gullnaz, Rebecca, Drake, & Josh, [podcast], YouTube.
44. Trust and Safety Professional Association, (28 March 2023), From NGOs to T&S | Laura, Liz and Paola, [podcast], YouTube.
45. Shulruff, T. - Arizona State University, (2022), Trust & Safety, A Snapshot of the Field.
46. Trust and Safety analyst salary in the United State | Salary.com.
47. Trust and Safety Manager salary | Salary.com.
48. TSPA's recently completed global compensation survey for T&S professionals will provide further information on remuneration when reported.

## Chapter 2

49. Goode, L., (21 November 2023), Twitter's former head of trust and safety finally breaks her silence, WIRED.
50. Integrity Institute, Leadership Advice for New Trust and Safety Leaders.
51. 5Rights Foundation, (2023), Updated Disrupted Childhood: The cost of persuasive design.

52. Ibid.
53. Even when shareholders do call on companies to prioritise safety, they face significant challenges in getting motions passed when voting rights are structured to favour founders. See De Meulemeester, C., (1 June 2022), [ESG resolution round-up: Founders' voting power helps Meta and Amazon fend off ESG proposals](#), Responsible Investor.
54. Goode, L., (21 November 2023), [Twitter's former head of trust and safety finally breaks her silence](#), WIRED.
55. Lenhart, A., & Owens, K., [The Unseen Teen](#), Data & Society.
56. Milmo, D., (12 September 2024), [Parents 'don't use' parental controls on Facebook and Instagram, says Nick Clegg](#), The Guardian.
57. For example, YouTube removed [8.3 million](#) videos in the first quarter of 2024 for violating its Community Guidelines. It is estimated that [3.2 million](#) videos are uploaded to YouTube per day making the total number of videos for the same period approximately 288 million. This is a takedown rate of approximately 1.1%. Similarly, TikTok removed [1.6%](#) of videos during the same period.
58. Béjar, A., [We must make social media safe for teens](#).
59. (3 February 2022), [Facebook: Daily active users fall for first time in 18-year history](#), BBC. Notably, the first wave of layoffs that included T&S followed just months later.
60. Rushe, D., & Milmo, D., (3 February 2022), [Facebook's first ever drop in daily users prompts Meta shares to tumble](#), The Guardian.
61. [Meta Stock Price October 2021](#) | StatMuse Money.
62. Bonomo, E. B., (25 October 2021), [What the Facebook whistleblower did to the company's stock in 6 weeks](#), TIME.
63. Safety-related decisions to withdraw advertising spend appear to be influenced safety issues but primarily by loss of confidence in a platform as a reliable "brand custodian". See [More marketers to pull back on X \(Twitter\) ad spend than ever before](#).
64. Bonomo, E. B., (25 October 2021), [What the Facebook whistleblower did to the company's stock in 6 weeks](#), TIME.
65. In October 2024, Roblox's share price fell by 4% following serious allegations about risk of harm to children. The investment fund also claimed that Roblox over inflated its Daily Active Users metrics and growth and profit forecasts. This makes it difficult to know what caused the price to fall. See Saul, D., (8 October 2024), [Roblox inflated data and doesn't protect underage gamers, short seller Alleges—Stock falls 4%](#), Forbes.
66. Blodget, H., (23 August 2011), [Mark Zuckerberg on Innovation](#), Business Insider.
67. Johnson, B., (21 February 2017), [Privacy no longer a social norm, says Facebook founder](#), The Guardian.
68. In 2012, it issued a Consent Order prohibiting the company from misrepresenting the privacy or security of consumers' personal information, and the extent to which Facebook shared personal information with third parties. See Federal Trade Commission, (2011), [Decision and order](#). In 2019, the FTC fined Facebook \$5 billion for violating the terms of the 2012 order. See Federal Trade Commission, (24 July 2019), [FTC imposes \\$5 billion penalty and sweeping new privacy restrictions on Facebook](#).
69. Horwitz, J., (10 June 2019), [Apple's former top lawyer: \\$1 billion budget enabled high-risk strategies](#), VentureBeat.
70. Ibid.
71. Integrity Institute, [Focus on features | Prevent harm through design](#).
72. Béjar, A., [Good questions for the press and others to ask social media companies](#).
73. For example, capacity within engineering teams (critical to building safety systems and products) is often stretched with competing demands on their time including requests from profit centres which may take priority over requests from T&S. See Allen, J., Crews, A., Louie, J., Motyl, M., Lawson, A., Gurley, S., (29 January 2024), [Child Online Safety Briefing](#), Integrity Institute.
74. UK Government's Department for Science Innovation and Technology, (October 2024), [Technology and Trust and Safety: the state of play and integration of technology](#).
75. Spectrum Labs, (2021), [Making the Case for Trust and Safety](#).
76. 5Rights Foundation, (2021), [Pathways: How digital design puts children at risk](#).
77. Ibid.



78. Spectrum Labs, [White Paper | Making a Trust and Safety Business Case](#).
79. UK Government's Department for Science Innovation and Technology, (October 2024), [Technology and Trust and Safety: the state of play and integration of technology](#).
80. Supreme Court of the State of New York, (2024), [Complaint – The People of the State of New York v. TikTok Inc.](#)
81. New Mexico Department of Justice, (2 October 2024), [Attorney General Raúl Torrez Files Unredacted Complaint Against Snapchat, Exposing Internal Messages that Snap Knowingly Contributed to Harm Amongst Children](#).
82. Kaplan, J., (07 January 2025), [More Speeches and Fewer Mistakes](#), Meta. It includes a recorded announcement by Mark Zuckerberg, Meta CEO.
83. Meta, (2024), [Community Standards Enforcement Report - Q3 2024 report](#).
84. Meta, (22 January 2025), [Community Standards Enforcement Report on Bullying and Harassment on Instagram](#).
85. Meta, (07 January 2025), [Community Standards Enforcement Report on Hate Speech on Instagram](#).
86. Meta, (26 June 2024), [Community Standards Enforcement Report on Graphic Violence on Instagram](#).
87. Ibid.
88. Isaac, M., Frenkel, S., & Conger, K., (10 January 2025), [Inside Mark Zuckerberg's Sprint to Remake Meta for the Trump Era](#), The New York Times.
89. Ibid.
90. See [the final notice imposed on UBS AG](#) by the Financial Services Authority in 2012.
91. Other common recommendations include comprehensive training for all team members (not just for those in oversight roles, a zero-tolerance approach to those who don't follow compliance procedures and continued, proactive efforts to surface and understand risk.
92. Trust & Safety Professional Association, [Approaches to Trust & Safety](#).
93. UK Government's Department for Science Innovation and Technology, (October 2024), [Technology and Trust and Safety: the state of play and integration of technology](#).
94. [Transcript of the evidence session of the Joint Committee on the draft Online Safety Bill | Frances Haugen](#), (25 October 2021)
95. [Transcript of the evidence session of the Joint Committee on the Draft Online Safety Bill](#), (28 October 2021).
96. Horwitz J., Blunt K., (22 December 2023), [A 'Recipe for Disaster': Insiders Warned Meta's Privacy Push Would Shield Child Predators](#), The Wall Street Journal.
97. Béjar, A., (07 November 2023), [Written Testimony of Arturo Béjar before the US Senate Judiciary Subcommittee on Privacy, Technology](#).
98. Béjar, A., (07 November 2023), [Written Testimony of Arturo Béjar before the US Senate Judiciary Subcommittee on Privacy, Technology](#). See also Béjar, A., (30 September 2024), [Feedback to the European Commission on the Protection of Minors Guidelines](#).  
For example,
  - did not collect data on unwanted sexual advances, despite internal research found that 13% of 13- to 15-year-olds experienced the issue in the past seven days;
  - claimed that “views of violating content that contains suicides and self-injury are very infrequent... many times we do not find enough violating samples to precisely estimate prevalence”, while its internal research found that 8.4% of 13- to 15-year-olds experienced the issue in the past seven days;
  - reported a prevalence rate of 0.2-0.3% for hate speech in its Transparency Report, despite internal data indicating that 26% of 13- to 15-year-olds experienced the issue in the past seven days.
99. Ibid.
100. [Transcript of the evidence session of the Joint Committee on the Draft Online Safety Bill](#), (28 October 2021).
101. Béjar, A., (07 November 2023), [Written Testimony of Arturo Béjar before the US Senate Judiciary Subcommittee on Privacy, Technology](#). See also Béjar, A., (30 September 2024), [Feedback to the European Commission on the Protection of Minors Guidelines](#). Additionally, a shareholder initiative to require Meta to publish data on harms to children in its annual report was blocked in line with the recommendation from the company's Board, Meta, (2024), [Proxy Statement to shareholders](#).



102. Meta, (2024), [Proxy Statement to shareholders](#).
103. YouTube, [YouTube Community Guidelines enforcement](#), in Google Transparency Report.
104. TikTok, (26 September 2024), [Community Guidelines](#), in Enforcement Report.
105. X, [Global Transparency Report H1 2024](#).
106. Snapchat, (25 April 2024), [Transparency Report 1 July 2023 - 31 December 2023](#).
107. TikTok does report the global number of accounts removed for being suspected underage.
108. Integrity Institute, [Focus on Features - A project of the Integrity Institute](#).
109. Lexology, (20 July 2021), [What is Age-Appropriate Design? What Does It Mean To Be Age-Appropriate When Designing Online Content?](#)
110. Goode, L., (21 November 2023), [Twitter's former Head of Trust and Safety finally breaks her silence](#), WIRED.
111. UK Government's Department for Science Innovation and Technology, (October 2024), [Technology and Trust and Safety: the state of play and integration of technology](#).
112. Trust & Safety Professional Association, [Key elements of a Trust & Safety team](#).
113. UK Government's Department for Science Innovation and Technology, (October 2024), [Technology and Trust and Safety: the state of play and integration of technology](#).
114. Ibid.
115. Center for Humane Technology, (07 June 2024), [Former OpenAI Engineer William Saunders on Silence, Safety, and the Right to Warn](#), [podcast], YouTube. See also William Saunders' [written testimony](#) to the US Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law for the Hearing on Oversight of AI: Insiders' Perspectives, (17 September 2024).
116. Open letter, (signed 4 June 2024), [A Right to Warn about Advanced Artificial Intelligence](#).
117. Quote provided for this report.
118. [Safety is Sexy](#), [podcast], in Incident Management with Eric Han, (03 September 2024).
119. Integrity Institute, [Leadership Advice for New Trust and Safety Leaders](#).
120. Carville, O., D'Anastasio C., (22 July 2024), [Roblox Predator Problem Potentially Exposes Kids to Pedophiles](#), Bloomberg.
121. Free Press, (2024), [Big Tech Backslide: How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections](#).
122. X's written response to Senate questions in March 2024 included the following "Today, we have approximately 2300 people working on Trust and Safety matters". This suggests a cut of 43% since Musk bought the company.
123. eSafety Commissioner, (1 November 2024), [Report reveals the extent of deep cuts to safety staff and gaps in Twitter/X's measures to tackle online hate](#).
124. See [post on X from EU's former Commissioner for Justice](#), Didier Reynders, on 24 November 2022. The European Commission followed up his comment with a statement emphasising that "*risk management obligations also include a strong component on the appropriateness of the resources allocated to managing societal risks in the Union. Among other matters, the Commission will scrutinise the appropriateness of the expertise and resources allocated, as well as the way they organise their compliance function.*" Source: Lomas N., (24 November 2022), [Twitter layoffs trigger oversight risk warning from Brussels](#), TechCrunch.
125. Goode, L., (21 November 2023), [Twitter's former Head of Trust and Safety finally breaks finally breaks her silence](#), WIRED.
126. Elliott, V., (06 November 2023), [Big Tech Ditched Trust and Safety. Now Startups Are Selling It Back As a Service](#), WIRED. See also Corral, C., Stringer, A., (01 May 2024), [A Comprehensive Archive of 2023 Tech Layoffs](#), Tech Crunch. Elliott, V., (03 May 2023), [Twitter Really Is Worse Than Ever](#), WIRED. Goggin, B. (29 March 2024), [Big Tech companies reveal trust and safety cuts in disclosures to Senate Judiciary Committee](#), NBC. (20 February 2025), [TikTok restructures trust and safety team, lays off staff in unit](#), sources say, Reuters.
127. Field, H., Vanian, J., (26 May 2023), [Tech layoffs ravage the teams that fight online misinformation and hate speech](#), CNBC.
128. For example, Věra Jourová, the EU's vice-president in charge of compliance with the code on disinformation said of the Twitter redundancies: "*I am concerned about the news of firing such a vast amount*

- of staff of Twitter in Europe. If you want to effectively detect and take action against disinformation and propaganda, this requires resources". See also Lomas, N., (24 November 2022), [Twitter layoffs trigger oversight risk warning from Brussels](#), TechCrunch.*
129. Shulruff, T. - Arizona State University, (March 2024), [Trust & Safety in Everyday Technologies](#).
  130. McIntyre, N., (28 November 2023), [Bumble, Grindr, and Hinge moderators struggle to keep users – and themselves – safe](#), WIRED.
  131. Allyn, B., (12 May 2020), [In settlement, Facebook to pay \\$52 million to content moderators with PTSD](#), NPR.
  132. Wiessner, D., (13 July 2022), [YouTube settles moderators' case over graphic videos for \\$4.3 mln](#), Reuters.
  133. Hatmaker, T., (24 March 2022), [Former TikTok content moderators file lawsuit over 'psychological trauma'](#), TechCrunch.
  134. Rowe, N., (2 August 2023), ['It's destroyed me completely': Kenyan moderators decry toll of training of AI models](#), The Guardian.
  135. Hellerstein, E., (28 February 2023), [Content moderator lawsuits against Big Tech pile up](#), Coda Story.
  136. Shulruff, T. - Arizona State University, (March 2024), [Trust & Safety in Everyday Technologies](#).
  137. Samra, R., (2018), [Brief history of burnout](#), BMJ.
  138. Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., DeMarco, J., (2023), [Content Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional Study](#), *Cyberpsychol Behav Soc Networking*, 27(2), 149-155.
  139. Williamson, V., Murphy, D., Phelps, A., Forbes, D., Greenberg, N., (June 2021), [Moral injury: the effect on mental health and implications for treatment](#), *The Lancet Psychiatry*, 8(6), 453-455.
  140. Williamson, V., Murphy, D., Greenberg, N., (17 July 2020), [COVID-19 and experiences of moral injury in front-line key workers](#), *Occupational Medicine*, 70(5), 317-319.
  141. Ibid.
  142. Integrity Institute, [Leadership Advice for New Trust and Safety Leaders](#).
  143. [Revealing Reality, Cleaning up in the attention economy](#).
  144. Shulruff, T. - Arizona State University, (March 2024), [Trust & Safety in Everyday Technologies](#), p. 2.
  145. Shulruff, T. - Arizona State University, (March 2024), [Trust & Safety in Everyday Technologies](#), p. 5.
  146. Open letter, (signed 4 June 2024), [A Right to Warn about Advanced Artificial Intelligence](#).
  147. Trust and Safety Professional Association (TSPA), [The Purpose and Role of T&S Teams](#).
  148. Bhuiyan, J., (9 October 2021), ['Welcome to the party': five past tech whistleblowers on the pitfalls of speaking out](#), The Guardian.
  149. For transparency, the author declares she participated in an employee restricted share scheme whilst employed at TikTok.
  150. Examples of tech companies that offer employee share schemes (or equivalent for companies that have not yet gone public) include Alphabet, Meta, Bytedance (which owns TikTok), Roblox, Snap, AirBnB, Deliveroo, Amazon and Apple.
  151. [Transcript of the deposition of Anika Collier Navaroli before the US House of Representatives Select Committee to investigate the January 6th attack on the US Capitol](#), (1 September 2022).
  152. For example, in the US, the [American Institute for Chartered Professional Accountants](#) and the [Chartered Institute of Management Accountants](#).
  153. For example, in the US, the [Royal Institute of Chartered Surveyors](#).
  154. For example, in the UK, the [Solicitors Regulatory Authority](#).
  155. For example, in the US, the [Communication Workers of America](#).
  156. For example, in the US, the [American Medical Association](#).
  157. For example, in the US, the [National Air Traffic Controllers Association](#).
  158. For example, in the US, the [National Education Association](#).
  159. A list of resources and organisations is set out in the Further Reading section.
  160. Integrity Institute, (8 February 2024), [How We Helped with the Senate Hearing on Child Safety Online](#).
  161. [Comments posted on LinkedIn following TrustCon24](#) include "T&S professionals are often few in number at any single organization, and simply being surrounded by other folks who work in these spaces felt like a weight off my shoulders." "Deeply grateful for Trust & Safety Professional Association for our yearly collective catharsis".

- 162. Solicitors Regulation Authority, [SRA Standards and Regulations](#).
- 163. General Medical Council, [What is Good medical practice?](#).
- 164. National Association of Social Workers, [Code of Ethics: English](#).
- 165. UK College of Policing, [Code of Ethics](#).
- 166. [Aviators Model Code of Conduct](#).
- 167. UK Council of Psychotherapy, [UKCP Code of Ethics and Professional Practice](#).
- 168. International Federation of Accountants, [Code of Ethics for Professional Accountants](#).
- 169. National Society of Professional Engineers, [NSPE Code of Ethics for Engineers](#).

### Chapter 3

- 170. Allen, J., Hunsberger, A., (12 February 2025), [The Future of Trust & Safety: Navigating Challenges in a Shifting Industry](#), [podcast], Integrity Institute.
- 171. Integrity Institute, (19 January 2024), [Child Safety Online](#).
- 172. Including express provisions on T&S such as minimum resourcing requirements (including for startups), clarity on autonomy and authority and protected status within governance frameworks so that advice must be taken into account.
- 173. For example, the European Commission's Digital Services Act, Ireland's Online Safety and Media Regulations and the UK's Online Safety Act.
- 174. Allen, J., Crews, A., Louie, J., Motyl, M., Lawson, A., Gurley, S., (29 January 2024), [Child Online Safety Briefing](#), Integrity Institute.
- 175. Allen, J., Hunsberger, A., (12 February 2025), [The Future of Trust & Safety: Navigating Challenges in a Shifting Industry](#), [podcast], Integrity Institute.
- 176. Ofcom, [Consultation at a glance: our proposals and who they apply to](#).
- 177. [DSA whistleblower tool](#), in Digital Services Act, European Commission.
- 178. Allen, J., Hunsberger, A., (12 February 2025), [The Future of Trust & Safety: Navigating Challenges in a Shifting Industry](#), [podcast], Integrity Institute.
- 179. [Transcript of the deposition of Anika Collier Navaroli before the US House of Representatives Select Committee to investigate the January 6th attack on the US Capitol](#), (1 September 2022).
- 180. Donig, D., (20 November 2024), [The path forward for responsible tech — All tech is human](#), All Tech Is Human.
- 181. In the global south the security of a content moderator's salary may offer security for wider family members (even when it is substantially lower than peers in major hubs such as San Francisco, Singapore and Dublin) making it harder to walk away from jobs even when they have a negative impact on mental health and wellbeing.
- 182. Spence, R., Bifulco, A., Bradbury, P., & Martellozzo, E., (2023), [The psychological impacts of content moderation on content moderators: A qualitative study](#), Cyberpsychology Journal of Psychosocial Research on Cyberspace.
- 183. Bhuiyan, J., (9 October 2021), ['Welcome to the party': five past tech whistleblowers on the pitfalls of speaking out](#), The Guardian.
- 184. Efforts by T&S professionals to support legislative proposals in the US indicate they do not believe the regulatory system of oversight of tech is sufficient to keep users and wider society safe.

### T&S Assessment Tool

- 185. Open letter, (signed 4 June 2024), [A Right to Warn about Advanced Artificial Intelligence](#)
- 186. Mental health is better supported by giving T&S operatives providing authority, autonomy and agency to pursue safety objectives than by offering wellbeing products such as VR headsets and apps, or employee wellness programmes. Monitoring of mental health through surveillance of T&S's operatives activities creates privacy risks that far outweigh any protective effects.

