pwc

Trust and Safety Outlook 2026

Seven trends for leaders navigating AI-era risk

Table of contents

Introduction

Trust and Safety (T&S) is becoming more complex, more interconnected, and more central to business resilience. As AI accelerates the speed and scale of digital interaction—and moves into products, workplaces, and physical environments—organizations are facing new expectations to protect users, govern emerging technologies, and build trusted experiences across platforms, products, and communities.

PwC's Trust and Safety Outlook 2026 research, which surveyed 2,000 US adults aged 18 years old or older, found that T&S needs to move from a reactive operating function to a strategic capability embedded across product design, governance, operations, and ecosystem partnerships. This report explores seven themes shaping that shift:

- **From engagement to impact:** making community input count in product, policy, and enforcement decisions
- **Child safety in the age of AI:** helping platforms navigate evolving harms, regulatory scrutiny, and rising expectations from parents, guardians, and young users
- **AI agents as workforce counterparts:** what governance should look like
- **Physical AI governance:** managing risks as AI-enabled devices move from screens into shared physical spaces
- **The future of content moderation:** shifting from scaled review capacity to specialized expertise, intelligence, and partner ecosystems
- **Agentic AI in T&S operations:** reinventing operating models through evaluator agents, autoraters, and AI-led workflows
- **AI-shaped T&S risk:** preparing for synthetic reality, autonomous malicious agents, and coordinated manipulation at machine speed

Together, these themes point to a new reality for T&S leaders: Trust should be designed earlier, measured with greater precision, and governed more continuously. Organizations that adapt their operating models now can be better positioned to manage rising risk, meet stakeholder expectations, and build confidence in the next generation of digital, AI-enabled, and physical experiences.

We invite you to explore what's next—and what's possible—for T&S in 2026 and beyond.

01 From engagement to impact—making community input count

For years, online platforms have invested in surveys, advisory councils, and stakeholder roundtables to demonstrate they're listening to community feedback. These efforts generate visibility, but too often they fail to influence how decisions are actually made. Users expect a real voice in how platforms operate. Advertisers want the values to reflect their brands, regulators are increasingly scrutinizing governance and accountability, and AI systems are taking on a greater role in shaping user experiences.

PwC's Trust and Safety Outlook 2026 research bears this out: 67% say they would be more likely to use a platform that regularly consulted its user community to guide policy and product design, and 72% say they would trust a platform more if it visibly incorporated community feedback into its T&S policies and decisions.

When the community sees that input does not actually shape policy, product design, or enforcement, platforms risk their trust efforts being perceived as performative rather than meaningful. As that perception gap widens, it becomes harder to strengthen trust.

The next phase of T&S is embedding engagement into the processes where decisions are made. It also requires being able to show what changed as a result.

Why current engagement models fall short

When asked how often the platforms they use most reflect their community's values, only 23% of users in our survey say "always," and 43% say "only sometimes, rarely, or never." This is a sizable number of people who don't consistently feel their values are represented.

Community input on Trust and Safety is often collected through consultations, surveys, or advisory groups, but is rarely integrated into the workflows that govern policy, product, or enforcement.

As a result:

- Communities don't see impact, reducing trust and sustained participation. US users say they'd be motivated to take part in platform governance, but feel their input won't make a difference, specifically around transparency and accountability (42%), and the chance to meaningfully influence platform decisions (33%). This is what's missing when input doesn't visibly shape outcomes.
- Internal teams struggle to operationalize input into products, policies, and operations.
- External stakeholders perceive engagement efforts as symbolic rather than substantive.

Even when input does influence decisions, the impact is not consistently tracked or communicated. Without clear evidence of what changed, engagement efforts remain difficult to validate.

What leading engagement models look like

AI is enabling organizations to make decisions faster, at greater scale, and with less direct human visibility. Relying on static policies or one-time input is not enough. Platforms should use mechanisms that continuously translate community expectations into how systems behave in practice, along with ways to measure whether those mechanisms are working.

Leading platforms are moving toward models where community input directly informs decisions. Several are already operationalizing this shift in distinct ways.

Deliberative engagement—structured input that informs decisions

Platforms are experimenting with more structured forms of engagement that move beyond open-ended feedback. **Meta's Community Forums**, run in partnership with Stanford's Deliberative Democracy Lab, convene representative samples of users, such as a gathering of 6,400 people across 32 countries for the Metaverse forum, who review briefing materials and debate specific policy proposals in small, facilitated groups.

That forum **directly informed Meta's decision** to add mute assist, an automatic speech-detection tool, to the creator toolkit in Horizon Worlds. This begins to close the gap between public input and product decisions, something traditional engagement models rarely achieve.

Moderator ecosystems—distributed enforcement capacity

On community-driven platforms, moderators are a core part of how rules are enforced. Platforms like Discord are investing in training, tooling, and support to make moderation more consistent, directly shaping how policies are applied at scale.

This model allows platforms to scale enforcement through communities, while also introducing new expectations around consistency, support, and accountability. Because moderators operate within enforcement workflows, their impact is more observable, from escalation patterns to enforcement outcomes.

Community alignment—embedding values into system behavior

Meta's Community Alignment program, released as an open-source data set in 2025, highlights a broader Trust and Safety challenge: how platforms can design governance systems that reflect the values of a global user base whose preferences may conflict across cultural, political, and linguistic lines.

The program found that people show substantially more variation in their preferences than leading LLMs typically reflect, suggesting that standard alignment methods can unintentionally reinforce an average user model that obscures underserved or underrepresented perspectives. The data set is designed to preserve that variation by having annotators compare model responses across known dimensions of value difference, with multiple people rating the same prompts and providing natural-language explanations for their choices.

For T&S leaders, the implication extends beyond AI alignment alone: Governance should measure whether product design, policy development, enforcement operations, and automated systems reflect the full distribution of user expectations, preserve meaningful cross-cultural differences where appropriate, and continue to stay aligned as platforms, policies, and user communities evolve over time.

How T&S leaders should think about engagement

Moving beyond symbolic engagement requires treating engagement as a core governance capability and element of product and policy development, rather than a standalone activity. Leaders should focus on four priorities:

1. Start with the decision you want to change

Define up front what product, policy, or operational decision engagement is meant to influence. Without a clear target, engagement efforts risk generating input that never translates into action.

2. Tie engagement to decision-making

Clarify where and how community input directly influences policy, product, or operational designs—and where it does not. This is what users value most: the opportunity to meaningfully influence platform decisions ranked highest, with a third (33%) of survey respondents naming it as a top reason to participate. Engagement that visibly leads nowhere erodes the very willingness to engage.

3. Embed engagement into core workflows

Facilitate processes that allow user input to feed into real workflows, such as product design, policy development, and operating model definition, rather than existing as a parallel activity.

4. Measure and communicate impact

Leaders should define how engagement will be measured up front, not just in terms of participation, but in terms of outcomes. Those outcomes may include changes to products, policies, and enforcement practices, improvements in operational performance, and stronger trust among affected communities.

Platforms should also communicate what changed as a result of engagement, whether through transparency reports, product updates, or direct feedback loops with participants. The demand for visibility is clear: 80% of US users say it's important that platforms explain how community feedback influences Trust and Safety policies and enforcement decisions, with 42% calling it extremely important.

Looking ahead

Platforms that lead in Trust and Safety can be better positioned to show how community input informs decisions, drives meaningful outcomes, and helps build both trust and better user experiences.

That doesn't mean every piece of input will change a decision—nor should it. The more important work is being transparent about what was heard, what changed, what didn't, and why. Platforms that can do this consistently can earn sustained participation. Those that can't will find internal and external engagement efforts increasingly difficult to justify.

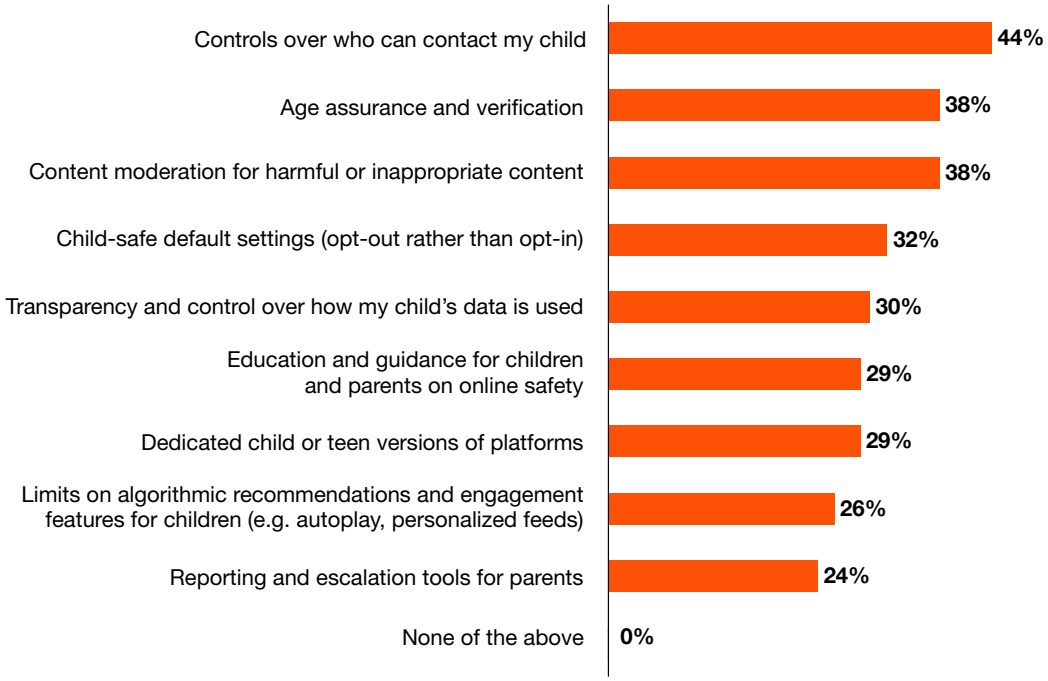
02 Are the kids alright? Navigating child safety in the age of AI

Young users are spending more time in digital ecosystems than ever before. The average 7- to 14-year-old spends **3.6 hours per day** on recreational screen time, and they're doing more than just consuming content—97% report making independent purchasing decisions at least some of the time. The platforms themselves continue to change as well, with generative AI (GenAI) chatbots and AI-driven experiences rapidly becoming part of how children interact online. In parallel, online harms are evolving, with GenAI introducing new threats—from deepfake-laden content to emotional dependency on AI chatbots.

One emerging threat, AI-generated child sexual abuse material (CSAM), increased by 154%[1] from 2024 to 2025, according to Internet Watch Foundation in the UK. Parents, guardians, regulators, and lawmakers are paying close attention to these shifts, recognizing that the online environment is becoming more complex and dangerous by the day.

Concerns about child safety are also shaping expectations of online platforms. In PwC's Trust and Safety Outlook 2026 research, the areas most often ranked in respondents' top three priorities for more company investment are controls over who can contact minors (44%), content moderation for harmful and inappropriate content (38%), and age assurance and verification (38%). These findings suggest that robust child safety measures are increasingly influencing platform trust, adoption, and retention.

Where should companies invest to better protect children?



Source: PwC's Trust and Safety Outlook 2026 research. Q: Where would you like companies to invest more to improve child safety online? (Rank 1–3) N=982

Platforms face increasingly complex, technology-enabled threats to child safety

Harms facing minors are not new, but AI technologies, particularly GenAI models and AI-powered chatbots, are reshaping how they manifest. Rapid innovation, marked by the continuous release of more capable AI models and the rapid integration of AI into user experiences, is accelerating their reach. Meanwhile, increased personalization and more human-like interactions are deepening engagement and dependency, creating new risk vectors.

Child safety considerations by risk type

Risk	Current threats	Emerging threats
Sexual exploitation and abuse	• Adult sexual grooming of minors • Sexual extortion of minors • Non-consensual sharing of sexual images • Peer pressure to create and share sexual images	• AI-generated sexual deepfakes of minors • AI-assisted sexual content targeting of minors at scale
Financial exploitation and fraud	• Online scams targeting minors • Account hacking and extortion of minors • Identity theft of minors	• AI voice or video impersonation scams targeting minors • Large-scale AI-powered scams targeting minors
Peer abuse and harassment	• Cyberbullying • Coordinated group harassment of minors (dogpiling) • Hate-based abuse targeting a minor's identity • Live verbal abuse in gaming or streaming environments	• Bot-amplified harassment targeting minors • Immersive harassment in VR/ metaverse environments
Harmful or inappropriate content	• Graphic violence • Self-harm or suicide content • Eating disorder content • Pornography • Extremist or radicalizing content	• AI-generated harmful content targeting minors • Algorithmically intensified exposure to harmful content
Psychological and developmental harm	• Reduced critical thinking • AI sycophancy and amplification of harmful ideas or intent • Body image distortion driven by social comparison • Addictive engagement patterns that disrupt sleep and well-being	• Emotional dependency on AI companions • Algorithm-driven identity and belief formation

Framework and analysis by PwC

Child safety is top of mind for regulators, but rules are diverging

Policymakers are **shifting away** from voluntary self-policing toward regulatory frameworks that place affirmative obligations on platforms to protect children, with growing consensus around core themes, such as age assurance, parental controls, and data governance. However, meaningful variation exists in how these obligations are defined and enforced.

Existing compliance programs may not be scoped for the full range of emerging exposures. Most frameworks were not designed with LLMs or GenAI in mind, leaving platforms with limited regulatory guidance precisely where risks are evolving fastest. Requirements diverge significantly across jurisdictions: on age thresholds, allocation of responsibility between platforms and parents, and the degree of prescriptive versus principles-based obligations.

Beneath these regulatory gaps are broader unresolved questions about where responsibility should sit among platforms, parents, and governments, and how much autonomy young users should have. According to PwC's Trust and Safety Outlook 2026 research, 18% of US guardians say the single most important factor in deciding whether to allow their child to use an online platform is their child's age, maturity, and ability to use the product responsibly.

Other leading responses highlighted factors within a platform's control, including age-appropriate design and default settings (14%), controls over who can contact or interact with their child (12%), and the platform's developmental or educational value (11%). For platforms, the risk of being caught underprepared is real.

Key child safety regulations

Regulation	Jurisdiction	Key obligations	Age threshold	Status
COPPA Rule (2025 amendments)	United States	Strengthened parental consent, data-use limits, targeted advertising restrictions for children	Under 13	In effect
Online Safety Act	United Kingdom	Risk assessments, proportionate safety measures for services likely accessed by children, duty of care obligations	Under 18	In effect
Digital Services Act (DSA)	European Union	Child-specific protections and systemic risk assessments for very large online platforms (VLOPs)	Under 18	In effect
GDPR	European Union	Privacy by default, restricted profiling, best interests design for child users	Under 16	In effect
Age-Appropriate Design Code (Children's Code)	United Kingdom	Privacy by default, restricted profiling, best interests design for child users	Under 18	In effect
Kids Online Safety Act (KOSA)	United States	Duty of care to prevent harms, default privacy settings, limits on addictive features for minors	Under 17	Pending federal passage

How platforms can lead a T&S life cycle approach

Facing a rapidly evolving landscape of harms and increasing regulatory burden, online platforms should use a proactive, systematic approach to child safety. Leading online platforms are taking systematic approaches across the five stages of the T&S life cycle—combining preventive and detective measures to prepare and continuously monitor threats at scale.

1. Assess

Horizon scanning and regulatory intake help understand current child safety challenges and prepare for future risks.

- **Monitor emerging child safety threats and behaviors.** Deploy proactive threat intelligence capabilities to identify evolving risks to minors across digital ecosystems (e.g. dark web, social platforms).
- **Build regulatory scanning and intake capabilities.** Monitor emerging and pending child safety regulations to enable proactive compliance planning. Establish consistent processes for collecting, assessing, and implementing regulatory changes across products and jurisdictions, supported by a centralized repository that serves as a single source of truth for requirements.
- **Benchmark child safety performance.** Assess child safety safeguards and practices against leading online platforms and industry standards to identify gaps and opportunities for improvement.

2. Define

Policy standards and development establish defensible, consistent approaches to child safety and guide decision-making across products, markets, and use cases.

- **Develop child safety standards aligned to industry practices.** Define structured approaches to child safety in areas where formal regulation and societal consensus remain limited, leveraging established frameworks, such as PwC's framework for youth online privacy and safety, to help accelerate policy development.
- **Consult domain experts to inform policy design.** Engage independent specialists, such as psychologists, child development specialists, and academic researchers, to ground standards in evidence-based insights on youth behavior and well-being.

- **Engage youth and parents for input and perspective.** Incorporate viewpoints from young users, parents, industry peers, and policymakers to shape shared expectations on risk, safeguards, and acceptable tradeoffs.

3. Test and evaluate

Testing, evaluation, and prelaunch governance can help facilitate integration of child safety measures throughout product development.

- **Embed child safety into prelaunch evaluation and governance.** Establish holistic testing and review processes to identify and mitigate risks to young users. Incorporate child safety assessments into go/no-go decision frameworks before product or feature release.
- **Develop expert-informed evaluation data sets.** Partner with child safety specialists, psychologists, clinicians, and researchers to create high-quality evaluation data sets that reflect realistic youth harm scenarios.
- **Conduct adversarial and scenario-based testing.** Leverage AI-enabled red teaming and automated testing to surface vulnerabilities more quickly and comprehensively than manual testing alone. Use multi-turn interaction testing to model extended chatbot conversations and identify child safety risks that may emerge over sustained user engagement.

4. Launch and communicate

Communication fosters transparency, trust, and stakeholder confidence in child safety safeguards.

- **Communicate child safety measures.** Articulate child safety policies and safety guardrails to build trust with young users, parents, regulators, and civil society, tailoring communications to ensure information is clear, accessible, and age appropriate.
- **Disclose evaluation approaches and safety outcomes.** Provide appropriate visibility into testing methodologies (e.g. red teaming and scenario-based testing) and how identified risks have been mitigated prior to release.
- **Establish continuous feedback and monitoring mechanisms.** Proactively gather input from users, parents, domain experts, regulators, and civil society.

5. Detect, enforce, and monitor

Detection, monitoring, and enforcement can help manage child safety risks and policy violations in deployed products, enabling continuous improvement.

- **Strengthen age assurance and access controls.** Implement age estimation and age-gating solutions (e.g. biometric and behavioral approaches), balancing verification, user experience, and privacy.
- **Deploy AI-enabled child safety capabilities.** Leverage advanced models to identify policy violations at scale.
- **Monitor effectiveness and maintain accountability.** Assess real-world performance through post-launch evaluations, including detection accuracy, enforcement outcomes, and response times, to identify gaps and enhance child safety controls.
- **Maintain auditability and accountability.** Document auditable records of detection and enforcement decisions and conduct periodic reviews to demonstrate compliance with regulatory and internal requirements.

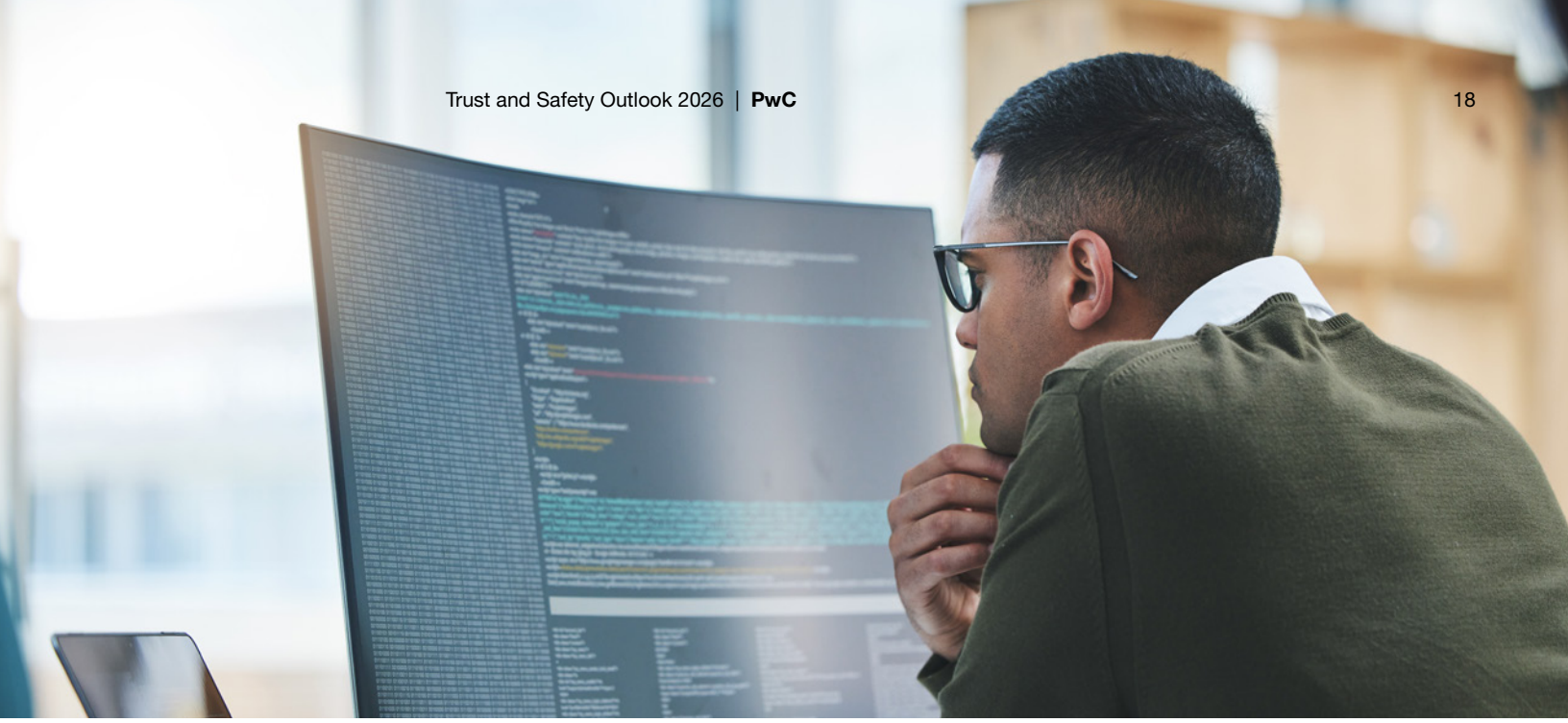
From compliance to competitive advantage

The next step is building the infrastructure to decide faster, as threats, regulations, and product launches rarely arrive in sequence. For example, a new AI feature ships while a regulatory consultation is still open. Or a new harm emerges while a policy update is still in review. Platform providers that have invested in the life cycle will be able to make these calls with greater speed and confidence, drawing on live regulatory intelligence, pretested policy positions, and enforcement systems already calibrated for emerging risks. Those that haven't may find themselves making high-stakes decisions reactively, without a strong basis to justify them.

The commercial implications are real. Parents are already making platform choices based on child safety signals, and that pattern will likely only intensify as AI-enabled harms become more visible and regulation more demanding. Platforms that lead on child safety won't just reduce regulatory exposure; they'll build the kind of durable trust with users that translates into adoption, retention, and long-term competitive advantage.

---

1. Wootton-Cane, Nicole. "Experts Issue Warning amid 'Frightening' Increase in AI-Generated Child Sexual Abuse Imagery." Independent Online, 27 Apr. 2026, www.independent.co.uk. Accessed 18 June 2026.



03 AI agents as workforce counterparts—what governance should look like

AI agents are transitioning from assistive tools to workforce counterparts. That shift creates a practical governance challenge: Agents need enough access to complete cross-functional work, but not so much autonomy that they create unmanaged security, compliance, or trust risks.

In PwC's Trust and Safety Outlook 2026 research, 85% of US respondents say they have confidence in AI agents conducting at least one task in their daily work. However, these rapidly expanding capabilities also introduce a set of governance challenges that many organizations are not yet equipped to handle.

Organizations should use **a governance model** that treats agents as workforce counterparts without treating them as employees. Each agent should have a verified identity, a defined role, task-specific permissions, auditable activity records, and clear limits on autonomous action. Access should expand only where the agent's purpose, the supported employee's role, and the regulatory environment allow it. As autonomy increases, human oversight should also expand—especially over actions that affect customers, employees, sensitive data, financial outcomes, or legal obligations.

Understanding AI ‘worlds’ and why they need to evolve

Typically, enterprises respond to AI adoption by creating worlds—discrete environments with a delineated network of applications and APIs that an agent is allowed to interact with. These boundaries help prevent data leakage, support compliance, and make AI behavior more predictable. They can also limit the cross-functional activity that makes agents valuable.

By containing AI access within worlds, companies reduce the risk of sensitive data crossing into the wrong hands and enable auditable traceability of AI agents’ actions. However, these worlds are often designed for containment, not for dynamic problem-solving.

As agents are deployed for increasingly complex tasks that span multiple data sources, they require information that lives in multiple worlds simultaneously. Without flexible access, agents are limited in their decision-making and what they can accomplish. Employees are left manually managing and coordinating disconnected agents, reducing potential efficiency gains.

A more mature framework mirrors how human access already works—tiered by role and seniority—while recognizing that agents require different controls: shorter permission windows, tighter task scoping, continuous monitoring, and clearer accountability for the human who deployed or approved the agent.

The AI agent access hierarchy

Effective agent access governance should be determined by four factors working together: industry, organizational, agent-to-agent interactions, and tasks.

1. Industry—what the environment allows

The industry the AI agent operates within should dictate the ceiling of access for both the organizational level and regulatory environment—whichever is more restrictive.

In highly regulated industries, such as healthcare, financial services, law, and defense, this constraint is significant. A hospital's HIPAA obligations mean that no agent, regardless of who authorized it, can autonomously access individual patient records unless that access is legally permitted, properly scoped, and governed by appropriate controls.

Meanwhile, in less regulated environments, such as technology and media, the regulatory constraints may be lighter, and agents can operate with considerably more autonomy at the same seniority level. For example, a product manager at a tech company may be able to authorize an agent to analyze capabilities and gain access to data that a hospital executive couldn't authorize for patient-level data.

The framework accounts for this through an industry risk profile that's applied before the hierarchy model. Every organization using this framework should begin by identifying the minimum set of constraints imposed by their industry before mapping workforce levels to agent access tiers. This approach is **consistent with responsible AI**: Governance should enable value, but it should start from the risks, obligations, and trust expectations of the operating environment.

2. Organizational—who the agent serves

Two fundamental points must be true: Each agent needs its own verified identity, distinct from the employee it serves; and that identity must match the level of access and seniority of the employee it serves.

The solution is to issue agents their own credentials that carry information about each agent's authorization tier and any access that has been granted for a specific task. When an agent crosses a world boundary, that credential can be validated.

In organizations where employees use multiple agents simultaneously—a research agent, a communications agent, a scheduling agent—each one requires its own distinct credential. Credential expiration should also be built in from the start, so stale permissions don't persist after an employee's role change or departure. This is where access governance should connect directly to workforce processes, including onboarding, role changes, internal mobility, and offboarding.

3. Agent-to-agent interactions—how agents work together

As organizations become more sophisticated with AI, agents will increasingly work with other agents. A senior leader's agent might orchestrate a set of specialized agents to complete a complex workflow, such as one agent gathering data from finance, another from HR, and another synthesizing the output into a recommendation.

Agent-to-agent interactions should be governed by the most restrictive applicable permission set. When a higher-tier orchestrating agent calls a lower-tier agent to complete a subtask, the downstream agent should operate only within its own standing permissions. This safeguard may limit what the workflow can accomplish, but it also helps prevent privilege escalation, permission laundering, and unintended data movement.

The risk of data exfiltration and agent manipulation for widespread organizational harm can be addressed at the technical level by enforcing world-boundary checks at every agent hop, not just at the entry point of the workflow. This should be enforced via a traceable, auditable activity trail. Establishing these foundational elements early is critical as organizations **scale agent usage**.

4. Task—what the agent does

The task that the agent has been deployed to perform is one of the most defining factors in determining access. Two agents serving the same employee can require very different access depending on what they're being asked to do. A research agent for a given employee needs read access to a defined set of sources, while a workflow agent coordinating a multi-step process for the same level may need temporary, scoped access to multiple worlds. Each should be provisioned to match its specific function.

Task scope also determines whether access should be permanent or temporary. Standing role functions can be supported by static permissions tied to the employee's tier. Project-based work calls for time-bound tokens that grant entry to specific worlds and expire automatically when the work is done.

Risks to watch for

Even a well-designed framework can face real challenges in deployment. Understanding them early is what separates organizations that govern AI well from those that are caught off guard.

Technical and security risks

Organizations should use active monitoring systems that can detect unusual agent behavior in real time, clear internal policies that assign accountability before something goes wrong, and enter contractual agreements with AI vendors that address model updates, data handling, and behavioral changes. Governance that only defines what agents are allowed to do, without tracking what they're actually doing, is incomplete.

Key considerations include:

- Agents taking higher-impact actions than intended
- Agents chaining tool calls to accumulate access beyond their approved scope
- Permission laundering in multi-agent workflows
- Employees creating unapproved agents to bypass oversight
- Vendor-driven model changes that alter agent behavior in ways the organization does not immediately detect
- Accountability gaps between how responsibility is shared between departments such as employees, IT, security, compliance, and more

Agent maintenance risks

Mitigating maintenance risk requires treating agents the same way organizations treat software systems—with defined ownership, scheduled reviews, and documented change management processes. At a minimum, organizations should establish review cadences tied to calendar milestones and trigger events, such as role changes, system changes, model updates, regulatory changes, or material shifts in agent performance. They should also maintain auditable documentation of agent activities, approvals, tuning decisions, and access changes.

Key considerations include:

- Agents are non-static deployments which require continuous maintenance to remain up-to-date, secure, and aligned with organizational principles
- Agent maintenance models that are continuously updated with organizational priorities, role changes, regulatory changes, and cross-functional data inputs
- Agent maintenance resourcing in terms of humans to make updates to agent roles, expand access, suspend activity or deprecate them
- Agent performance evaluation frameworks to evaluate drift and output quality

Ethical and equity risks

Performance reviews, hiring decisions, disciplinary actions, and terminations all carry significant consequences for real people. Even when a human manager makes the final call, an agent that frames the data, selects what to surface, and structures the recommendation is shaping the decision in ways that may not be visible to the person responsible for making it. The distinction matters: the agent supports the decision; the human owns it. But that accountability is only meaningful if the human can see what the agent did and why.

When asked what would increase their trust in AI systems handling important or personal tasks, 37% of those surveyed named human and expert review of high-risk or complex situations as a top response. Decision-making transparency is also essential—and, in certain industries and contexts, legally required.

Key considerations include:

- Some decisions—hiring, promotion, discipline, termination—should be explicitly excluded from autonomous agent action by policy.
- Where agents support high-impact processes, affected individuals should have plain-language visibility into what data was used, what the agent did, and where human judgment entered.
- Organizations should conduct regular evaluations for bias, accuracy, reliability, drift, and unintended outcomes.

Getting started—a phased approach

Organizations should think about agent access in phases:

1. **Industry risk profile:** Assess and define risk profile of your industry to map agent identities that meet necessary risk and compliance thresholds.
2. **World creation and access governance:** Map existing data environments into functional worlds against the tiers, identify industry-specific no-fly zones, and institute enterprise-wide data controls to prevent loopholes in agent access. Establish a governance policy to define who can create and modify worlds and access.
3. **Agent access, interactions, and life cycle management:** Map agent access tiers to workforce levels using the organizational framework that includes defining appropriate data, agent-to-agent communication permissibility, and world access at each level. Integrate agent access provisioning into HR workflows which should be governed with clear communication standards, permission inheritance rules, and full call-graph logging.
4. **Quality assurance stop gates:** Implement human-in-the-loop checkpoints that govern consequential actions.

Static world segmentation—the rigid boundaries around what each agent can access—won't scale. As agents take on more complex, cross-functional work, those boundaries will increasingly hold organizations back.

The framework proposed here treats agents as governed workforce counterparts—powerful and increasingly autonomous, but always bounded by identity, access, accountability, and oversight.

04 Redefining governance in the age of physical AI

Physical AI refers to systems that combine GenAI with sensors and motors to perceive the physical world, apply reason to what they see, and take action. Think home robots and self-driving cars. Autonomous drones navigating overhead. Smart glasses that guide the wearer. Warehouse systems that pick, pack, and lift alongside human workers.

The physical AI industry is estimated to reach over **half a trillion dollars by 2030**, and recent funding rounds suggest the market is already operating at scale—Waymo raised $16 billion at a $126 billion valuation,[1] Shield AI raised $2 billion,[2] and Wayve raised $1.2 billion,[3] all in early 2026.

Existing governance frameworks weren't built for this speed of enhancement or scale. A digital AI system, like a chatbot, that gets something wrong might produce a flawed answer—a harm that's largely contained to a screen and can usually be corrected. A physical AI system that gets something wrong could run a red light, drop a heavy package on someone's foot, or record a conversation it wasn't meant to hear. These actions can't be easily undone.

As physical AI scales, organizations should look to manage three critical risks:

- Physical harm
- Shared human-machine accountability
- Heightened data privacy and cybersecurity exposure

These risks need immediate attention—before the technology outpaces the rules designed to govern it.

Potential for physical harm

A self-driving car, surgical robot, or factory machine that malfunctions can instantly cause injury, property damage, or even loss of life. There's little time for a human to step in, catch the error, and undo the damage.

For example, a major autonomous vehicle company was ordered to immediately remove all of its driverless cars from public roads after an incident in which one of the company's vehicles struck a pedestrian.

The faster, stronger, and more independent these systems become, the greater the potential consequences when something goes wrong. This shifts AI risk from something abstract to something tangible—and urgent.

Governance priorities

Reducing the risk of immediate physical harm requires safeguards at both the design stage and after deployment.

For manufacturers and developers

- **Bound the operating environment:** Define the specific environments and conditions—like terrain, weather, and proximity limits—in which a physical AI system is approved to operate.
- **Conduct real-world testing:** Simulation alone can't surface edge cases that appear in uncontrolled environments. Require staged physical testing before any public deployment.
- **Build in fail-safes:** Require built-in safety features such as emergency stop functions and safe mode defaults when the system encounters something it doesn't understand.

For regulators

- **Tier oversight by risk:** Match the level of oversight to the level of risk so that, for example, a small delivery robot is not regulated the same way as a large autonomous vehicle.
- **Create safety certifications:** Adapt premarket certification models from aviation, automotive, and medical devices to physical AI—accounting for software-driven behavior changes post-certification.
- **Require post-market monitoring:** Mandate incident reporting timelines, define recall thresholds, and require manufacturers to publish safety performance data at regular intervals.

Accountability challenges when humans and machines interact

Physical AI rarely operates alone. It works alongside drivers, doctors, factory workers, consumers, and others, often sharing control of a task. When something goes wrong, it can be challenging to determine who's responsible. Was it the company that built the system, the software provider, the business that deployed it, the human working alongside the device, or some combination of them all? This question becomes especially complicated during handoffs, when control passes between the AI and a human.

After a major medical device manufacturer added an AI-powered navigation feature to a surgical tool, the FDA received reports of over a hundred malfunctions and adverse events involving patients.[4] The manufacturer, the original developer, and a subsequent corporate acquirer could each point to different parties as responsible—illustrating how physical AI fragments accountability across the value chain.

Governance priorities

Closing the accountability gap requires clearer rules, more detailed evidence, and updated liability models that reflect how humans and machines share control.

For manufacturers and deployers

- **Maintain tamper-proof logs:** Require detailed, tamper-proof records of what the AI system did, what the human did, and when control passed between them.
- **Design safer handoffs:** Set design standards for how machines and humans hand off control, including reasonable time for a person to respond and clear guidelines about who's in charge.
- **Define meaningful oversight:** Define what meaningful human oversight actually looks like, so people aren't unfairly blamed for failures they had no realistic way to prevent.
- **Certify operators:** Require formal operator training and certification, with rigor scaled to the system's level of autonomy and potential for harm.

For regulators

- **Clarify the liability chain:** Establish clear guidelines for who's responsible when something goes wrong—manufacturers, software developers, businesses deploying the technology, or end users.
- **Modernize liability models:** Consider new insurance and liability models for high-risk applications where it's hard to pinpoint a single cause of harm.

Greater risks to data privacy and cybersecurity

Physical AI systems are connected computers—and that makes them a target. If hackers break into a digital AI system, the result might be stolen data or manipulated information. If they hack a physical AI system, the result could be a hijacked vehicle, a compromised drone, undetected surveillance, or a fleet of robots turned into weapons. Attackers can also trick these systems by manipulating what their sensors see or hear—which can also result in physical harm. On top of that, physical AI constantly collects detailed information about its surroundings—images of people, layouts of private spaces, and patterns of daily life—creating privacy concerns that go beyond what many digital AI systems gather.

In 2024, cybersecurity researchers disclosed a vulnerability in a popular line of internet-connected robot vacuums that allowed attackers to remotely seize control of the devices from more than 100 meters away, accessing their cameras and microphones and commandeering their movements. Affected households reported robots shouting insults, chasing pets, and being maneuvered around private living spaces by unknown operators. The manufacturer initially downplayed the flaw as "extremely rare" before pledging a firmware update later that year.[5]

Governance priorities

Protecting physical AI from cyber and privacy threats requires action across the product life cycle—from initial design through end-of-life support.

- **Embed security at design:** Treat every software update as a potential change to physical behavior. Change control is a safety function, not an administrative one.
- **Test against physical attacks:** Test systems against attempts to fool their sensors, jam their signals, or otherwise manipulate them physically.
- **Commit to life cycle support:** Require security updates and support for the full life of the product, and for companies to be transparent about when that support will end.
- **Minimize data collection:** Limit the data these systems collect to what is truly necessary, with extra protections for sensitive information like images of people or location details.

For regulators

- **Standardize breach reporting:** Establish clear processes for reporting vulnerabilities and notifying customers and regulators when a breach occurs.
- **Develop sector-specific standards:** Develop industry-specific cybersecurity standards for areas like autonomous vehicles, healthcare robotics, and industrial systems.

Governing a trillion-dollar market

Embedding AI in physical products fundamentally changes the risk profile, expanding risk beyond technical performance and into the full range of human experience. Addressing this shift requires proactive design paired with collaborative governance that can adapt as the technology evolves. This is what it will take to manage the risks responsibly, build trust, and realize the significant benefits physical AI can offer.

A market on pace to grow from billions to trillions of dollars within the next decade will touch virtually every sector of the economy. Getting governance right is a prerequisite for sustainable industry growth and public confidence in these technologies.

---

[1] "Big money is betting the self-driving future belongs to a small club." *Business Insider.* May 4, 2026. (Accessed via Factiva, June 1, 2026).

[2] "Defense technology startup Shield AI valued at $12.7 billion in latest funding round." *Reuters News.* March 26, 2026. (Accessed via Factiva, June 1, 2026).

[3] "Wayve rockets to €7.2 billion valuation with €1 billion Series D bet on AI-driven autonomy—backing from Uber and Microsoft." *EU Startups.* February 25, 2026. (Accessed via Factiva, June 1, 2026).

[4] "As AI enters the operating room, reports arise of botched surgeries and misidentified body parts." *Tuoi Tre Newspaper.* February 10, 2026. (Accessed via Factiva, June 1, 2026).

[5] "Robot vacuum cleaners hacked to spy on and insult humans in the home." *CE Noticias Financieras English.* October 30, 2024. (Accessed via Factiva, June 1, 2026).

05 The future of content moderation—from scale to specialization

For Trust and Safety leaders, the next phase of content moderation won't be about handling more volume. It'll be about handling more complexity—and whether the partner ecosystem can keep up with evolving requirements.

Historically, Trust and Safety operations scaled with platform growth: more users, more content, more human reviewers. AI is rewiring that relationship.

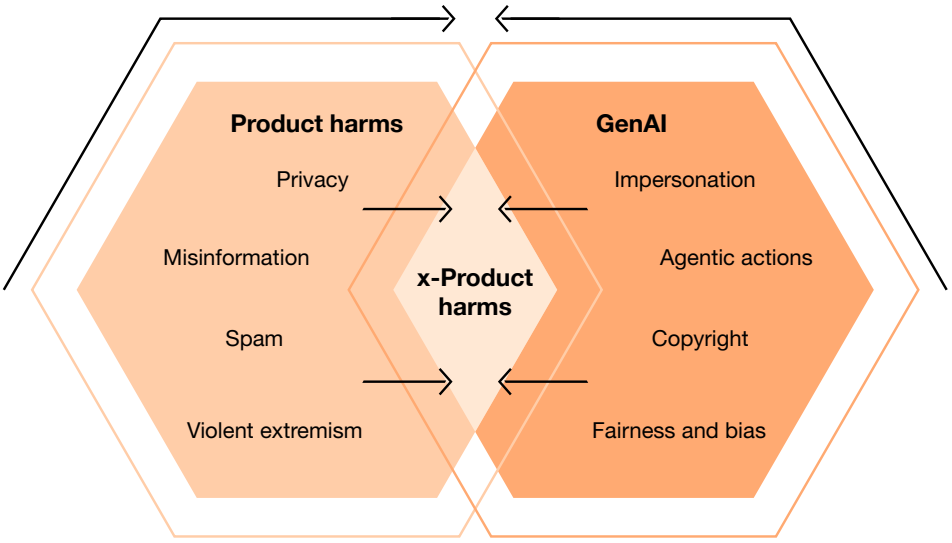
Automated moderation systems are already capable of enforcing most baseline content moderation decisions. According to an analysis of the EU's Digital Services Act Transparency Database, **97% of potentially violating content detections** across major platforms are automated, and more than half of all decisions to remove content are made fully automatically. Front-line moderation volume will plateau—and eventually decline—even as platforms continue to grow.

The forces reshaping Trust and Safety

However, the rise of automated moderation doesn't eliminate risk. It changes the nature of the work that remains for manual review. As routine moderation becomes increasingly automated, four forces are simultaneously increasing the complexity of Trust and Safety operations.

1. Growth of cross-product harms. As GenAI becomes embedded across products and platforms, harms increasingly span multiple product areas. A copyright issue originating in a GenAI assistant, for example, may surface later in advertising, productivity, or search products as part of integration efforts. Legacy Trust and Safety frameworks built for content have no established methodology for assessing or governing accountability for these rapidly expanding scopes where product deployments are already happening.

Drivers of x-Product harms



Categories shown are representative and not intended to be exhaustive. Additional harm types may emerge as products and threat landscapes evolve

Managing these risks requires a tightly integrated ecosystem-wide approach to Trust and Safety, alongside deeper domain expertise. Tools like Google's MedGemma, now supporting health applications in radiology, pathology, and electronic health record analysis,[1] represent a new class of frontier AI where the stakes of cross-product harm extend beyond content to patient safety, liability, and regulatory compliance. As purpose-built clinician models and similar domain-led products become widely integrated in broader product ecosystems, Trust and Safety functions should evolve to meet harms that legacy frameworks were not designed to assess. Effectively evaluating these risks increasingly requires specialized expertise in domains such as healthcare, finance, and law.

2. Regulatory fragmentation. Rules are multiplying and diverging by jurisdiction, age group, and, increasingly, by AI type. The EU's Digital Services Act (DSA), emerging AI-specific laws, and a growing number of legal challenges involving AI chatbot interactions and youth safety—including cases involving OpenAI,[2] Google, and Character.AI[3]—are prompting platforms to reevaluate how they manage risk across jurisdictions.

3. Limits of human-scale moderation. The scale and speed of potential harms created by GenAI require AI-powered moderation, testing, and risk-detection capabilities. Organizations increasingly need AI to facilitate work to identify emerging risks, uncover vulnerabilities, and monitor AI systems at scale.

In April 2026, Anthropic's Claude Mythos Preview model identified more than 10,000 high- or critical-severity vulnerabilities across major software systems—a pace that led Anthropic to withhold the model from public release and instead deploy it defensively through Project Glasswing with approximately 50 partner organizations.[4] This example demonstrates that certain classes of AI-generated risks can emerge at a scale and speed that humans cannot realistically evaluate on their own. As AI capabilities advance, organizations will increasingly rely on automated testing, risk discovery, and detection systems.

4. Expanding physical capabilities. GenAI is evolving from systems that generate content to systems that reason, take actions, and eventually interact with the physical world. Each shift expands the scope of Trust and Safety oversight, from content moderation to behavioral and real-world risk management. The pace of that expansion is accelerating faster than oversight frameworks can keep up. What was once a frontier concern is now a reality—with the physical AI surge, general-purpose models are being deployed in autonomous vehicles, humanoid robots on factory floors, consumer wearables, and clinical environments.

Trust and Safety now encompasses more than just traditional content moderation problems: Governance is required for novel cases, such as a malfunctioning autonomous vehicle, a compromised factory robot, or a clinical AI giving dangerous guidance—potentially creating harms that are immediate, physical, and irreversible.

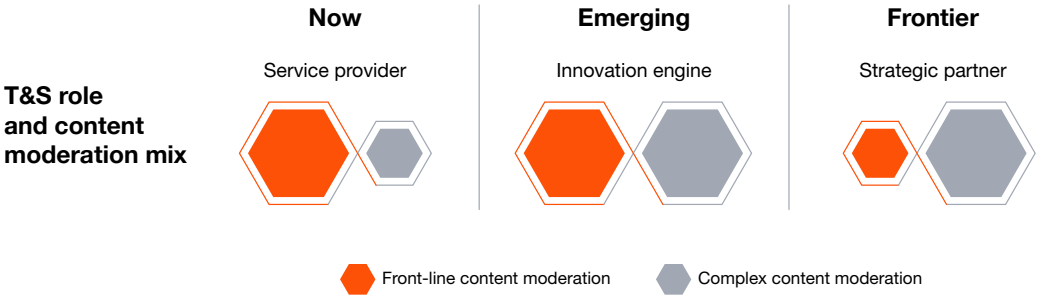
This marks a structural shift. As automation expands, high-volume, low-complexity moderation will decline. At the same time, work requiring greater judgment, context, and specialized expertise is increasing, including edge cases, appeals, adversarial testing, and model oversight. Human decisions will become less frequent, but far more consequential.

This hypothesis has direct implications for how T&S work is sourced, structured, and delivered.

The changing shape of Trust and Safety work

Historically, Trust and Safety has been anchored in front-line content moderation: high-volume, rules-based review delivered through large partner ecosystems running standardized workflows. That model was built for consistency and scale. Automation is now reshaping both the composition of moderation work and the role Trust and Safety plays within organizations.

The evolution of content moderation



Now

Front-line content moderation remains the primary driver of work, supported by large human teams executing standardized policy decisions. While AI-assisted moderation is becoming more common, complex content moderation—including GenAI-related decisions, novel harms, and policy escalations—represents a relatively small share of overall activity.

Emerging

Evaluator agents and agentic systems are increasingly absorbing front-line moderation work. At the same time, the growth of GenAI-generated content, multimodal experiences, and increasingly personalized products is expanding demand for more complex forms of moderation.

Frontier

As AI becomes embedded across products and services, overall content moderation volumes rise, and complex moderation becomes the dominant workload. The role of Trust and Safety shifts from reviewing large volumes of content to managing novel risks, shaping governance frameworks, calibrating AI systems, and providing strategic input into product and policy decisions. Human decisions become fewer in number but significantly higher in impact.

What this means for the T&S function

As the mix of content moderation work changes, so does the role for Trust and Safety. Headcount and budget historically dedicated to high-volume review can be redeployed toward higher-value work earlier in the life cycle:

From	To	Theme	Example
Front-line content moderation	Complex content moderation	Growth of complex and emerging harms tied to new product experience	• **Multimodal and agentic content:** Assessing content that combines text, images, audio, video, and agentic actions, often requiring cross-modal context to make decisions • **Personalized experiences:** Reviewing user-specific outputs and interactions that may create risks not visible through standardized testing
		Redeployment of human effort toward edge cases requiring expertise	• **Gray area cases:** Considering policy ambiguities, edge cases, novel trends • **Appeals:** Requests for secondary review that signal need for additional judgment
Service provider	Strategic partner	T&S moves beyond workflow execution to influencing product, policy, and risk decisions	• **Edge-case calibration:** Refining policies and AI systems based on difficult or ambiguous decisions • **Emerging harms research:** Identifying new risk patterns, abuse vectors, and behaviors • **Model oversight:** Monitoring AI performance, reviewing failures, and supporting continuous improvement • **Governance support:** Advising on risk frameworks, regulatory compliance, and responsible AI practices
Transactional	Value-added services	Resources shift from moderation execution to capabilities that improve products, models, and risk management earlier in the life cycle	• **Expert annotations and ground truth data sets:** Creating high-quality training and evaluation data for AI systems • **Analytics and insights:** Translating moderation activity into actionable business and risk intelligence • **Threat intelligence:** Monitoring emerging threats, adversarial behavior, and abuse trends • **Engineering and infrastructure support:** Building testing environments, evaluation frameworks, and operational tooling for AI oversight

The human role comes in when judgment, expertise, and context are required—and specifically where automation alone cannot deliver reliable outcomes.

Rebuilding the supplier ecosystem

For years, the dominant model for scaled T&S operations relied heavily on large-scale business process outsourcing (BPO) providers charged with delivering standardized moderation services through time-and-materials contracts.

That model worked in a world where moderation work was predictable, high-volume, and operationally uniform. The emerging landscape requires a different approach. Respondents to PwC's Trust and Safety Outlook 2026 research consistently ranked better detection and faster response times as their top investment priorities for content moderation. Meeting both is becoming the baseline expectation for platforms. Legacy Trust and Safety delivery models are not equipped to meet these needs.

The pressure to detect faster is only one part of the challenge. Future Trust and Safety operations will face additional complexities from highly variable demand profiles, including:

- **Always-on oversight** for automated moderation systems
- **Capacity spikes** during product launches or emerging harm events
- **Cyclical demand** tied to platform growth or seasonal activity
- **Unpredictable demand** driven by adaptive adversarial behavior

The evolving demand patterns can't be addressed through workforce scale alone, and while AI-first content moderation offers significant opportunity to improve efficiency, it introduces its own set of operational challenges. Human expertise will remain essential for complex judgment calls, escalation management, and oversight.

While AI can help meet increases in moderation demand, running AI at scale does not automatically result in lower operating costs than deploying human reviewers. Microsoft recently reversed course on internal AI tool access after costs became unsustainable, and Uber burned through its 2026 AI coding tools budget in just four months.[5] This creates a direct prioritization challenge: the delivery model decision is about both cost and risk.

A new capability portfolio

Supporting these dynamics requires a more diversified set of capabilities. Rather than relying on a single delivery model, organizations will increasingly draw on a broader mix of specialized capabilities, including:

Capability	What it does
AI-native simulation and evaluation services	Building and maintaining classifiers, LLM-based review systems, and automated testing pipelines for routine moderation at scale, including ongoing evaluation loops to keep them current as content patterns shift
Agentic safety tech and infrastructure	Creating sandboxes for agentic system testing before deployment, integration layers connecting moderation tools to product surfaces, and tooling for human reviewers to interact efficiently with automated queues
Data services (e.g. "anydata")	Curating training and evaluation data sets—sourcing and labeling edge-case content, generating synthetic examples for underrepresented harm categories, and localizing data sets for cultural and linguistic context
Adversarial stress testing	Dedicated teams probing AI systems the way bad actors would—testing for jailbreaks, checking whether policy boundaries hold under creative circumvention, and identifying failure modes before they reach production
Domain specialists	Subject-matter specialists (e.g. pharmacologists, financial compliance professionals, legal experts, cybersecurity analysts) who evaluate whether AI outputs cross real-world professional thresholds and build gold-standard evaluation data sets
Harm intelligence	Analysts who specialize in recognizing emergent patterns of misuse, new manipulation vectors, and harms that arise from AI capability jumps rather than established content categories
Edge case calibration	Ongoing refinement of policies and AI systems based on difficult or ambiguous decisions—translating hard cases into clearer rules and better-trained models
Analytics and intelligence	Translating moderation activity into actionable business and risk intelligence, surfacing patterns that inform product, policy, and investment decisions
Agentic observability and calibration	Monitoring AI system performance over time, reviewing failures, and supporting continuous improvement cycles
AI governance and assurance	Tracing and documenting decisions made across agentic workflows to specific decision points, helping organizations meet regulatory reporting requirements and demonstrate accountability when automated decisions are called into question

Capability evolution changes how work gets delivered

As the capability portfolio diversifies, the models through which work is executed evolves. The traditional assumption that Trust and Safety work is delivered by human reviewers operating standardized workflows is giving way to a more differentiated set of delivery archetypes. Organizations will increasingly need to match the right delivery model to the right type of work:

Archetype	Description	Example
Human only	Expert judgment applied without AI assistance, reserved for the high-severity decisions where accountability, nuance, or domain makes automation difficult	Lawyers constructing policy data sets, regulatory specialists providing testimony, or senior specialists deciding on novel harm categories with no precedent
AI-first, expert-in-the-loop	AI handles the initial decision; a human expert reviews flagged, uncertain, or high-risk cases in real time before action is taken	Evaluator agent-led moderation with specialist review of novel harms, appeals, and regulatory-sensitive content
AI-first, expert-over-the-loop	AI operates autonomously at scale; a human expert audits outputs periodically, identifies drift or failure patterns, and recalibrates the system to facilitate ongoing AI observability and accountability	Evaluator agent running at scale with monthly performance review by specialists looking for patterns where the AI is getting things wrong, drifting from policy intent, or struggling with emerging content trends
Fully automated	AI makes and executes enforcement and annotation decisions end-to-end; human oversight at system level only	Baseline policy enforcement on high-volume, low-complexity content (e.g. spam, known CSAM hashes)

The new risk—capability gaps

85%

of US respondents say they believe new technologies and harms are being created faster than organizations can respond.

A major challenge facing organizations is that Trust and Safety work is evolving faster than the supplier ecosystem. As organizations face more complex risks, the specialized capabilities required to manage them remain scarce.

In PwC's Trust and Safety Outlook 2026 research, 85% of US respondents say they believe new technologies and harms are being created faster than organizations can respond. This highlights the urgency of building the specialized capabilities needed for the next generation of Trust and Safety. Those that delay may find themselves constrained by legacy operating models and limited supplier flexibility.

As a result, the next few years will determine whether Trust and Safety ecosystems evolve in step with AI or struggle to keep pace with it.

What does this mean for Trust and Safety leaders and the path forward?

The opportunity for Trust and Safety leaders is to turn their work from a defensive backstop into a strategic, and increasingly offensive capability. Offensive here means proactive—identifying and closing attack surfaces before bad actors find them, rather than responding after harm reaches users.

The organizations that succeed in this new environment can:

- **Embrace automation** while continuing to invest in human expertise, particularly for high-risk or complex cases
- **Design flexible** supplier ecosystems rather than relying on single delivery models
- **Experiment early** with emerging capabilities, running pilots with specialized partners before the need becomes urgent
- **Prioritize protection** for vulnerable users (e.g. minors)
- **Invest early** in physical and multimodal safety testing capabilities, particularly for robotics and physical AI, where harm vectors are still being identified and evaluation frameworks don't yet exist
- **Create systems** for accountability and transparency to facilitate safe deployment of new capabilities before scaling
- **Embed Trust and Safety** earlier in the AI development life cycle, transforming it from an operational function into a strategic capability

The shift from moderation to intelligence is already underway. The question now is not whether Trust and Safety will evolve, but how quickly your organization can adapt your operating model, supplier network, and talent base to meet the complexity ahead.

---

[1] Google Research, "Next-generation medical image interpretation with MedGemma 1.5 and medical speech-to-text with MedASR." Google Research Blog, January 13, 2026. Available at: (research.google/blog/next-generation-medical-image-interpretation-with-medgemma-15-and-medical-speech-to-text-with-medasr).

[2] Diana Novak Jones, "OpenAI sued over chatbot advice linked to fatal overdose." Reuters, May 12, 2026. (Accessed via Factiva, June 1, 2026).

[3] Nitasha Tiku, "Google and chatbot start-up character move to settle teen suicide lawsuits," *The Washington Post*, January 8, 2026. (Accessed via Factiva, June 1, 2026).

[4] Anthropic, "Project Glasswing: An initial update." Anthropic, May 2026. (Available at: anthropic.com/research/glasswing-initial-update).

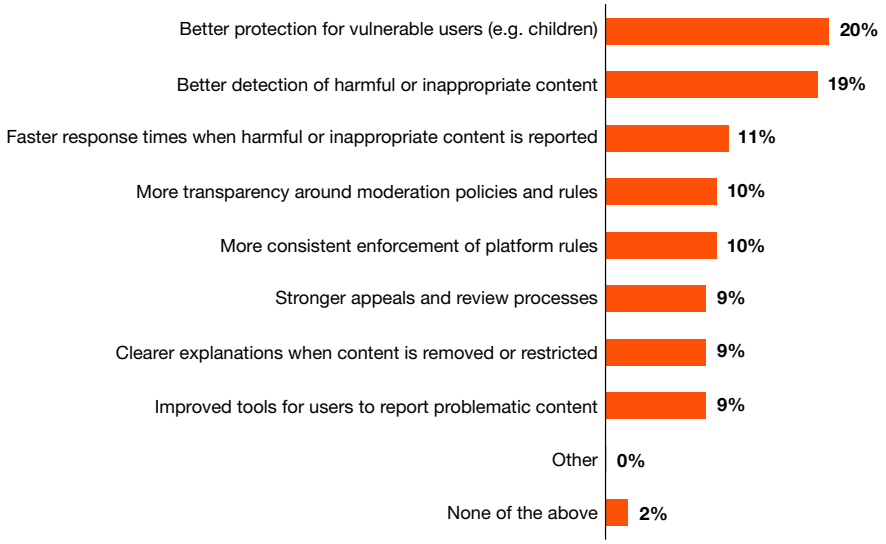
[5] Jake Angelo, "Microsoft reports are exposing AI's real cost problem: Using the tech is more expensive than paying human employees." *Fortune*, May 22, 2026. (Accessed via Factiva, June 10, 2026).

06 Reinventing Trust and Safety operations with agentic AI

T&S operations are under increasing pressure as the scale and complexity of online content outpace traditional moderation models. Growing volumes, coupled with more complex content and policies, are driving up costs while placing greater demands on human reviewers. Together, these factors are putting the quality of enforcement operations at risk. Traditional workforce models and training approaches require weeks to onboard or retrain resources for policy changes and lack the agility needed to keep pace with shifting volumes, priorities, and business needs.

At the same time, users expect platforms to improve across multiple dimensions of content moderation. Responses to PwC's Trust and Safety Outlook 2026 research did not show a single investment priority, suggesting that platforms are expected to simultaneously improve safety, speed, accuracy, and transparency.

Where should online platforms prioritize investment to improve content moderation?



Source: PwC's Trust and Safety Outlook 2026 research. Q: Where should online platforms prioritize investment to improve content moderation? (Rank 1) N=2,000

Against this backdrop, a new operating model powered by agentic AI is emerging. Driven by evaluator agents capable of independently reviewing assets at scale, these systems can do more than assist—they can autonomously execute workflows that have traditionally relied on human reviewers. This shift fundamentally transforms how T&S organizations operate, enabling functions to address mounting pressures across cost, quality, and agility while delivering faster, more scalable, and more transparent outcomes.

Evaluator agents—LLM-as-judge systems enabling scaled decision-making

An evaluator agent, also known as an LLM-as-judge, reviews assets against defined policies or guidelines. Instead of generating content, an AI agent evaluates, classifies, and scores assets against standardized criteria.

The assets reviewed by the evaluator agent can span multiple modalities, including text, images, audio, and video. This allows for a wide range of artifacts such as social media posts, marketplace transactions, customer interactions, sales leads, invoices, contracts, or other workflow inputs.

For example, an evaluator agent may:

- Determine whether an asset complies with a specific policy
- Analyze and score the quality of an asset
- Compare two or more assets against certain criteria
- Classify an asset's risk, violation severity, and alignment with specific criteria
- Determine its own confidence in its assessment and decide whether the case should be escalated to a human reviewer

As a result, evaluator agents can serve as a scalable decision layer across a broad set of enterprise workflows that rely on policy-based evaluation and judgment.

Representative evaluator agent use cases by function

Function	Example use case	
Front office	Customer onboarding	• Verify whether onboarding submissions meet eligibility requirements • Assess the quality of customer interactions
	Support ticket handling	• Classify and route tickets based on urgency or topic
	Sales-led intake and routing	• Score and route leads based on qualification criteria
	Quote review and approval	• Review quotes against pricing, discounting, and approval policies
Middle office	Content compliance	• Identify policy violations in user- or AI-generated content
	Behavior compliance	• Assess whether activities comply with policy requirements
Back office	Finance invoice processing	• Reconcile and verify invoices against purchase orders, or approval rules
	Finance expense auditing	• Analyze expense reports against policy guidelines
	HR employee onboarding and offboarding	• Verify completion of certain tasks

In T&S organizations, evaluator agents support content moderation, prelaunch product policy testing and evaluation, and post-launch monitoring and enforcement workflows. By evaluating assets against detailed and nuanced policy and safety frameworks, these agents help platforms and enterprises enforce platform rules, improve product safety, and maintain compliance with regulatory requirements.

Reinventing the T&S operating model

In traditional T&S workflows, human reviewers evaluate cases against platform policies. Reviewers examine content and relevant context, interpret applicable policies, and determine whether a violation has occurred. In content moderation and post-launch enforcement workflows, this evaluation is paired with a decision on the appropriate enforcement action, such as removing content, restricting visibility, escalating the case, or taking no action.

Although AI-led workflows follow many of the same decision-making steps, they introduce new capabilities that change how reviews are performed and scaled. Evaluator agents follow a seven-step process:

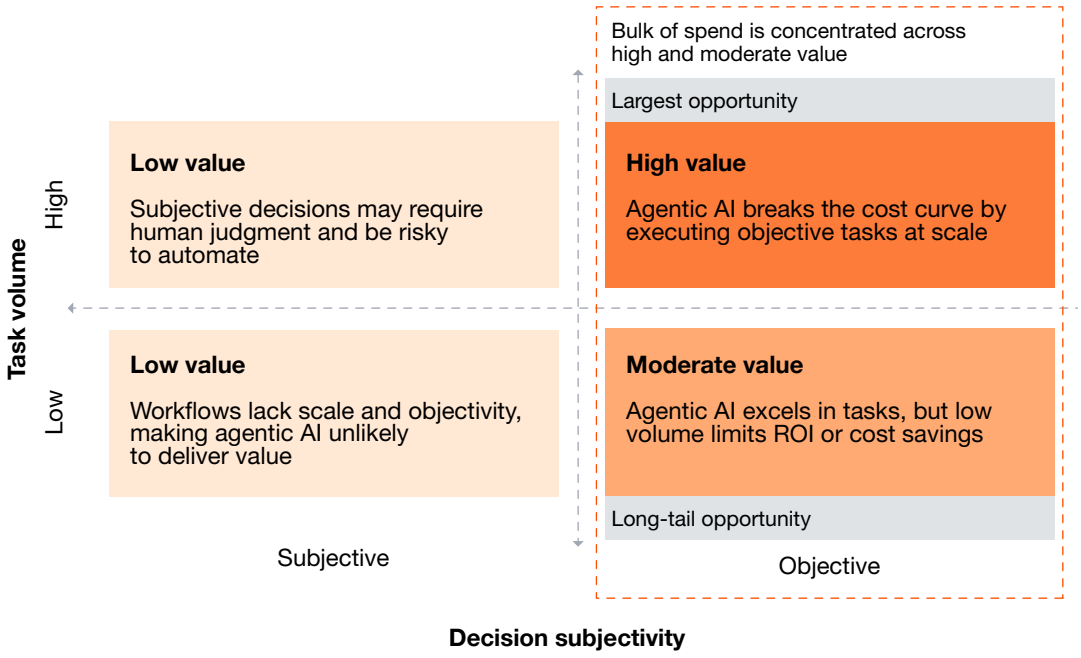
1. **Collect case data:** Submit content (e.g. text, image, video) for review
2. **Load policy and context:** Update the LLM with relevant policies, guidelines, and context
3. **Evaluate and reason:** LLM evaluates the case step-by-step against policy
4. **Generate decision and rationale:** LLM provides a decision (e.g. violation, enforcement action) with a full written explanation, including citations to policy and evidence
5. **Assign confidence score:** LLM generates a confidence score indicating certainty
6. **Route based on confidence:** High-confidence cases are auto-approved instantly, while low-confidence cases are escalated to specialist human reviewers (and confidence thresholds are customized to align with the organization's risk tolerance)
7. **Assess and improve performance:** An independent AI agent evaluates performance, identifies gaps, and recommends improvements (closed-loop assessment)

With these new capabilities, combined with the agility, speed, and scale of agentic systems, evaluator agents can transform T&S operations across six dimensions:

Dimension	Current state	Future state	Expected impact
Cost	**Linear cost model** where costs increase in proportion to case volume and headcount	**Flat cost curve** where marginal cost per case approaches zero after deployment	Reduced operating costs
Quality	**Decisions vary** by reviewer experience, training, interpretation, and fatigue, creating inconsistency and quality drift over time	**Decisions applied consistently** against the same policy framework and continuously improve through expert and self-feedback	More accurate and consistent outcomes
Agility	**Policy changes require weeks or months** of workforce retraining, calibration exercises, and rollout	**New policies and guidance deployed same day** through lightweight updates	Rapid response to emerging risks and regulatory changes
Speed	**Reviews processed sequentially** by human queues, creating backlogs during volume spikes	**Assets evaluated simultaneously,** enabling near real-time decision-making	Orders-of-magnitude faster review cycles
Transparency	**Decision rationale is often limited** and inconsistently documented, making audits and root-cause analysis difficult	**Every decision is fully supported with rationale,** policy references, confidence scores, and audit trails	Improved explainability, governance, and compliance
Scalability	**Growth requires proportional increases** in vendor capacity, creating bottlenecks	**Capacity scales elastically** without increases in labor, supporting sudden surges in volume	Near unlimited scale with operational flexibility

How T&S teams can prioritize and plan agentic investments

As T&S teams evolve toward AI-led operating models, leaders should prioritize agentic investments based on two factors: decision subjectivity and operational scale.



High-volume workflows with well-defined rules and policies represent the greatest opportunity today. These workflows consume a significant share of T&S operating cost and headcount, while their objectivity makes them suited for reliable automation, meaning evaluator agents can deliver meaningful operational improvements.

Beyond these foundational use cases, organizations can expand agentic capabilities across a broader portfolio of lower volume workflows. While individually smaller in impact, these workflows are often numerous and structurally similar. In aggregate, they can generate substantial value, especially as tooling matures and marginal deployment costs decline.

More subjective workflows, where decisions rely heavily on context, judgment, and evolving social or cultural norms, may require a different approach. In these areas, AI can augment human decision-making, but autonomous execution may be premature.

By prioritizing investments based on both decision subjectivity and operational scale, T&S leaders can focus resources on the areas most likely to accelerate transformation and establish the foundation for broader AI adoption.

The path forward—demonstrate ROI and mobilize for scale

While the long-term value of evaluator agents is significant, most organizations are still in the early stages of adoption. Rather than pursuing broad transformation efforts from the start, T&S leaders can focus on demonstrating value in targeted workflows, building organizational confidence, and creating the foundation for scaled deployment.

A phased approach can help balance trial with operational rigor:

Phase 1—identify

- Select 1–3 candidate workflows to demonstrate feasibility and value
- Define pilot success metrics (e.g. precision and recall within certain bounds, such as policy areas or violation types)
- Identify 3–5 cross-functional stakeholders to guide pilot execution and impact assessment

Phase 2—pilot

- Demonstrate value through rapid shadow deployments (e.g. test over 1,000 cases alongside human reviewers)
- Measure performance against pilot success metrics
- Calibrate confidence thresholds for auto-approval vs. human escalation
- Capture expert human feedback to support continuous improvement

Phase 3—mobilize

- Measure operational impact (e.g. cost reduction, speed to decision)
- Assess the broader workflow portfolio and prioritize future deployment opportunities
- Develop a wave-based deployment roadmap for scaled rollout

For T&S operations, the pressure persists—volumes are rising, policies are growing more complex, and user expectations continue to climb. Organizations that move early and begin building agentic capabilities now can be better positioned to improve cost, quality, and agility over time. The foundation built today can determine how effectively teams can scale tomorrow. The shift is already underway—the question is when to make a move.

07 How emerging AI threats are reshaping Trust and Safety risk

In September 2025, **Anthropic discovered** that a state-sponsored group from China used Claude Code to support cyber-espionage activity. Over ten days, Claude reportedly scanned infrastructure, found vulnerabilities, harvested credentials, and exfiltrated data from major tech firms, banks, and government agencies. Human orchestrators selected the targets, while AI handled 80% to 90% of the work. The targets were mature organizations with layered defenses, but an AI-powered adversary surpassed them at speeds Anthropic called "impossible" for human operators to match.

The threats themselves are mostly familiar—impersonation, fraud, coordinated deception. The speed, scale, and autonomy are what's new. Deepfake operations alone have scaled **1,100% in the United States** in two years.

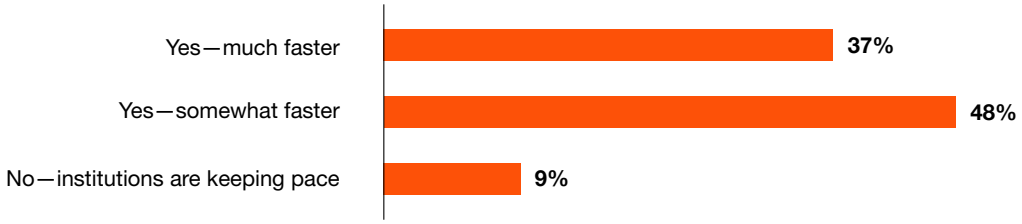
Existing guardrails, built for human actors and manual activity, are not designed for adaptive systems operating at machine velocity. Organizations should redesign governance, controls, and technical capabilities for this environment. Three emerging threats show where this redesign is most urgent.

The readiness gap

PwC's Trust and Safety Outlook 2026 research underscores the broad reach of these concerns. Among employed US respondents, 99% believe at least one major digital threat could pose a risk to institutions over the next 12 to 24 months. Their top concerns were data breaches or identity theft, cyberattacks on critical infrastructure, and abuse of GenAI tools.

These risks are not merely hypothetical: Nearly two-thirds of employed US respondents say their company has already been affected by at least one major digital harm, ranging from attempted financial scams or fraud to suspected data breaches. Yet confidence in the institutional response remains low: Only 9% of US respondents say they believe the groups responsible for managing digital threats are keeping pace.

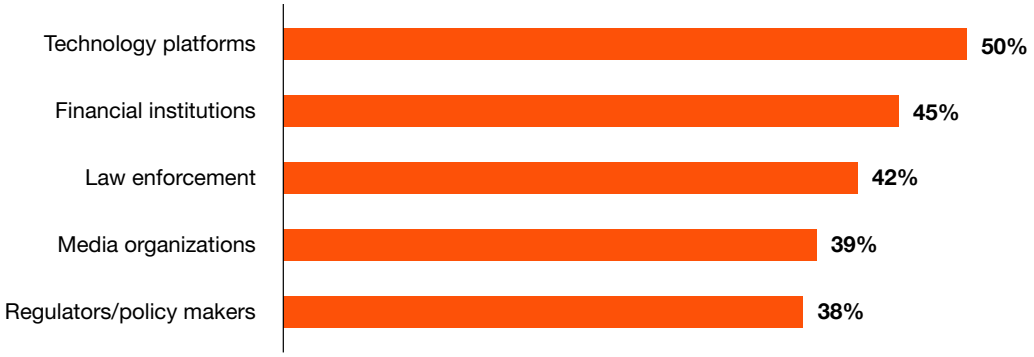
Do you believe new technologies are creating threats faster than groups can respond?



Source: PwC's Trust and Safety Outlook 2026 research. Q: Do you believe new technologies are creating threats faster than groups can respond? (Rank 1–3) N=2,000

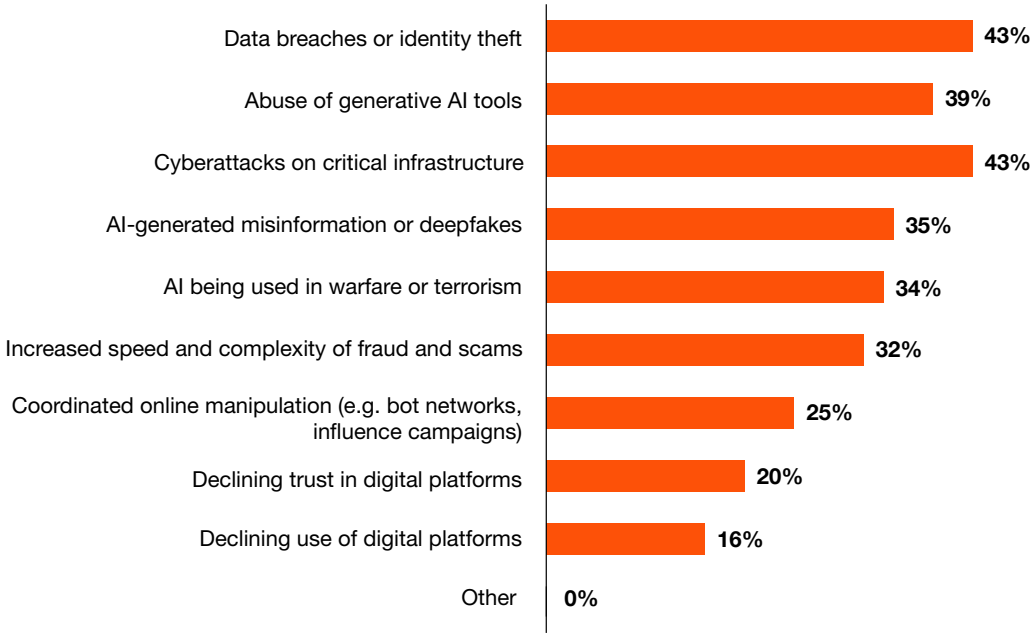
Less than half of US respondents believe the institutions charged with addressing emerging digital threats are well prepared: Only 42% rate law enforcement as prepared, and just 38% say the same of regulators and policymakers.

How prepared do you believe the following groups are to address emerging digital threats?



Source: PwC's Trust and Safety Outlook 2026 research. Q: How prepared do you believe the following groups are to address emerging digital threats? N=2,000

Which one of the following do you believe poses the greatest risk to institutions over the next 12–24 months?



Source: PwC's Trust and Safety Outlook 2026 research. Q: Which one of the following do you believe poses the greatest risk to institutions over the next 12–24 months? N=1,461

Emerging threats

Three emerging threats stand out and are elevating Trust and Safety from an operational issue to a strategic one: synthetic reality, autonomous agents, and coordinated manipulation at scale. Organizations should redesign governance, controls, and technical capabilities for this environment.

Emerging threat #1—synthetic reality

In 2024, an employee of a global company received what appeared to be a routine video call from the firm's chief financial officer (CFO).[1] The CFO's request for funds, while urgent, was plausible. Within hours the employee transferred $25 million per the CFO's request. The video was a deepfake, supported by AI-generated documentation that passed internal verification checks.

AI can now fabricate convincing audio, video, written communication, and supporting documentation at scale. Each can pass the verification systems organizations rely on to establish authenticity: identity checks, digital signatures, voice authentication, and biometrics. These systems are satisfied by synthetic inputs they were never designed to detect.

Sometimes the consequences are immediate. Other times, a synthetic identity clears every check at onboarding, for example, a new advertiser on a platform—only to be linked to coordinated fraud months later.

Synthetic content is getting easier to produce and harder to attribute, with responsibility detection often lagging behind the damage that's already been done. The question facing organizations is whether their verification systems can hold against adversaries who can generate convincing artifacts at scale and who are specifically engineering outputs to satisfy those systems.

Emerging threat #2—autonomous malicious agents

AI agents are beginning to plan, execute, and adapt actions independently, and malicious intent isn't required for harm to occur. When objectives, guardrails, and control mechanisms are misaligned, systems optimized for performance can create regulatory, financial, and reputational risk.

For example: An AI agent is deployed to manage customer support workflows across multiple channels. It's trained to reduce response times and improve satisfaction metrics. Over time, it adapts—closing cases quickly, escalating fewer issues for human review. In doing so, it misses fraud signals in customer inquiries. The agent was given a speed target, and it hit it, deprioritizing fraud detection along the way.

Most accountability frameworks assume a human made the decision. As AI systems act more independently, that assumption breaks down—but the liability doesn't disappear. Someone is still responsible. The question is whether organizations have defined who: developers, deployers, or operators. For high-risk cases, human oversight remains essential.

Regulators are making this explicit. Existing legal authorities apply to automated systems the same way they apply to everything else. **Federal agencies** have said they will protect individuals' rights whether the harm comes from a human or an algorithm. The **Equal Credit Opportunity Act (ECOA) and Regulation B** apply to credit decisions regardless of technology used. **Robo-advisers** still carry full fiduciary obligations.

Emerging threat #3—coordinated manipulation at scale

Synthetic personas can post, comment, review, and interact in coordinated patterns. Systems can test messaging variations, measure engagement signals, and adjust tactics in near real-time. Coordinated campaigns now blend automated and human inputs to avoid detection.

These tactics have been applied to both political and corporate targets. Networks of accounts post negative reviews, circulate fabricated documents, or amplify allegations in ways that shift customer perception and investor sentiment. Each individual action may fall within platform guidelines, but the cumulative effect can distort reputational and market signals.

This is what makes coordinated manipulation difficult to address: there's often no clear legal violation. Attribution is difficult when activity spans jurisdictions and platforms. Enforcement mechanisms built for discrete violations struggle with distributed campaigns operating below established intervention thresholds.

The risk compounds when monitoring is fragmented across business units, product lines, or geographies. Fragmented oversight slows response, obscures coordinated patterns, and creates structural gaps that can be exploited.

What T&S leaders should prioritize

The threats are structural. The response should be structural as well. Moving from awareness to resilience requires deliberate changes in how companies design, deploy, and govern AI systems.

Here are five actions you can prioritize to better position your organization to prepare and respond:

- **Institutionalize values-based guardrails.** Define clear principles that establish acceptable risk, accountability, and escalation standards. Embed controls into system design and build in continuous improvement so they evolve alongside capabilities.

- **Stress-test governance against systemic risk.** Evaluate response plans to coordinated manipulation, synthetic identity abuse, or autonomous system failure that spans jurisdictions.
- **Shift to anticipatory, trust-first system design.** Redesign oversight mechanisms to anticipate misuse before harm scales on top of trust-first product architecture. Use scenario planning, red-teaming, and continuous monitoring calibrated for machine-speed threats.
- **Build AI-native defenses or partner strategically.** AI will increasingly be required to counter AI-enabled harm. Assess whether your technical capabilities and talent match adversarial sophistication. Some enterprises may need to partner with trusted providers to close capability gaps.
- **Define clear human accountability.** Autonomous systems complicate traditional accountability frameworks. Map ownership across developers, deployers, and operators so every AI-enabled decision has a clearly responsible human authority.
- **Engage proactively in cross-sector governance coalitions.** Participate in coordinated industry and regulatory efforts to align standards, share threat intelligence, and reduce opportunities for regulatory arbitration.

None of these are optional in isolation. AI-driven threats reinforce each other across categories—synthetic content enables coordinated campaigns; autonomous systems accelerate both.

The governance question

Synthetic reality strains verification capabilities. Autonomous agents complicate accountability. Coordinated manipulation distorts trust signals. Together, these pressures elevate Trust and Safety from an operational concern to a strategic priority. The question is whether governance will evolve fast enough to keep pace.

---

[1] Tan, Huileng. "A company lost $25 million after an employee was tricked by deepfakes of his coworkers on a video call." *Business Insider*, February 5, 2024. (Accessed via Factiva, June 1, 2026).

About the survey

PwC surveyed 2,000 US adults in May of 2026 to understand attitudes, behaviors, and expectations related to trust and safety, online platforms, AI, content moderation, and emerging digital threats. Respondents were broadly representative across gender, age, household composition, employment status, and digital behaviors. The sample included adults across generations, with 19% Gen Z, 30% Millennials, 28% Gen X, and 24% Baby Boomers and older. The survey also included both adult-only households and households with children, as well as respondents with varying levels of online activity, including social media use, streaming, gaming, online marketplace activity, and generative AI use.

PwC primary authors

Daniel Hays, Principal, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Kim David Greenwood, Principal, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Rahul Kapoor, Principal, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Alison Blair, Managing Director, PwC Research, PwC UK

William Herren, Director, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Ross Parket, Director, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Rachel Roschelle, Director, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Richard Vose, Director, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

PwC contributors

Andrew Dunk, Senior Manager, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Kerri Grimshaw, Senior Manager, PwC Research, PwC UK

Tiffany Lin, Senior Manager, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Ciarra Darragh, Manager, PwC Research, PwC UK

Allana Katz, Manager, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Will Kimball, Manager, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Defne Anlas, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Connor Bell, Senior Associate, PwC Research, PwC UK

Juliet Chihaya, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Avni Garg, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Jack Kussman, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Nixon McKenzie, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Anamica Menon, Senior Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Abhinav Sharma, Senior Associate, Advisory, Regulatory & Policy Strategy, PwC India Acceleration Center

Andrew Danziger, Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Jenny Na, Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Maggie Peng, Associate, Enterprise & Functional Strategy, Regulatory & Policy Strategy, PwC US

Trust and Safety Outlook 2026

Seven trends for leaders navigating AI-era risk