Intouch CX™ | WebPurify An Intouch CX Company

E-BOOK

Top 13 Trust & Safety Trends for 2026: *Expert Predictions*

Introduction

From Australia's pioneering social media age bans to the rise of AI companion apps, from the explosion of synthetic content to watershed legislation protecting women from tech-facilitated violence: ***2026 promises to be a year of unprecedented transformation in how we protect people online.***

What began as isolated policy experiments are rapidly becoming global norms. Technologies that were once optional are becoming regulatory requirements. And the fundamental question of who holds power in content moderation decisions is being rewritten by new laws, new tools, and new user expectations.

But alongside progress come profound challenges. How do we prevent displacement from social media from creating a loneliness epidemic? How do we detect AI-generated fraud before it devastates vulnerable people? How do we ensure that the platforms connecting people in the physical world – from ride-sharing to gaming – can anticipate threats rather than merely respond to them?

We asked eight leading voices in Trust & Safety to share their predictions for the year ahead. Together, they identify 12 defining Trust & Safety trends for 2026 which paint a picture of an industry at a crossroads – one where the tools for change exist, but the will to deploy them universally remains the defining question.

The insights that follow come from senior leaders across policy, operations, safety research and regulation.

Featured Experts

Ailís Daly

HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

Ailís Daly leads trust and safety strategy across EMEA for WebPurify, an IntouchCX company. Ailís joined WebPurify as Head of Trust & Safety, EMEA, from TikTok, where she served as the Global Head of Violence & Aggression Issue Policies. With over 14 years in the tech industry, her work has spanned issues like online misogyny, human trafficking, international conflicts, and elections. She advises top-tier companies on regulatory requirements, Trust & Safety operations, and policy development across the region. She is also the host of Trust Issues, a Trust & Safety–themed podcast that shares insights from the people who keep the internet safe.

Anna MacCarthy Adams

SENIOR DIRECTOR OF CASE REVIEW, APPEALS CENTRE EUROPE

Anna MacCarthy Adams leads case review operations at Appeals Centre Europe, one of the EU's certified out-of-court dispute settlement bodies under the Digital Services Act. Previously, she served as EMEA lead for Revenue Policy at Twitter (now X), bringing deep expertise in content moderation appeals and user rights.

Alexandra Popken

SVP OF TRUST & SAFETY AND AI SERVICES, WEBPURIFY, AN INTOUCHCX COMPANY

Alexandra Popken is Senior Vice President of Trust & Safety and AI Services at WebPurify and a veteran in the field. Alex, WebPurify's first VP of Trust and Safety, joined from X (formerly Twitter), where she was the Head of Trust & Safety Operations. Alex's career highlights include building and scaling Twitter's trust and safety operations for monetization, growing it from a one-person team to a global organization.

Seyi Akiwowo

MULTI-AWARD WINNING FOUNDER AND AUTHOR OF 'HOW TO STAY SAFE ONLINE'

Seyi Akiwowo is a digital rights activist, founder of Glitch, and author focused on online safety, digital citizenship, and combating tech-facilitated abuse. Her work centers on empowering marginalized communities, particularly women and people of color, to navigate online spaces safely. She advocates for safety by design informed by public health approaches.

Dr. Rachel Kowert

FOUNDER OF PSYCHGEIST

Dr. Rachel Kowert is a research psychologist and consultant specializing in gaming communities, online behavior, and digital wellbeing. Through Psychgeist, she helps platforms understand the psychological dimensions of online safety and build evidence-based interventions. She is a strong proponent of proactive innovation and industry knowledge-sharing in trust and safety.

Emily Harman

TRUST & SAFETY EXPERT AND NCII ADVOCATE

Emily Harman is a trust and safety expert and former lead at OnlyFans, specializing in non-consensual intimate imagery (NCII) prevention and violence against women and girls (VAWG). She advises Stop NCII and advocates for mandatory hashing technologies, global regulatory harmonization, and maximized safety infrastructure over minimum viable compliance.

AJ Brown

COO, THE BRAND SAFETY INSTITUTE

AJ Brown is a leading voice in brand safety and media responsibility, helping organizations understand the business case for trust and safety investments. With expertise in risk mitigation, advertiser confidence, and ROI-driven safety strategies, AJ advocates for reframing trust nd safety as business enablement rather than merely compliance.

Sophie Walsh

DIRECTOR OF TRUST & SAFETY, DEPOP

Sophie Walsh is a senior leader in Trust, Risk, and Ethics with more than 18 years' experience leading international, multidisciplinary teams across the tech sector, online marketplaces, international NGOs, and the not-for-profit space. She is the co-author of multiple publications on risk management for the Institute of Risk Management and has served as a Trustee for an international NGO. Sophie also founded the International Good Practice Network, an initiative designed to help global NGOs collaborate more effectively to drive greater impact at scale.

Sarra Eddahiri

GLOBAL HEAD OF SAFETY, BOLT

Sarra Eddahiri leads global safety strategy for Bolt, one of Europe's leading ride-sharing and mobility platforms. Her expertise lies in real-world risk management, proactive threat detection, and building safety systems for platforms where digital decisions have immediate physical consequences. She champions anticipatory approaches over reactive pattern-based prevention.

1 Kids Will Be Banned From Social Media Around the Globe

"An early stage global normalization, a future where under sixteens may simply not have access to mainstream social media at all, or certainly not in the forum that we know it today."

Ailís Daly

HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

A dramatic shift coming in 2026, Ailís predicts, is the global acceleration of age restrictions on social media platforms. Australia is leading the charge, with its nationwide ban on users under 16 accessing social media platforms including Instagram, TikTok, Snapchat, YouTube and X that came into effect on December 10. Malaysia has announced a similar ban for 2026, while Denmark is pursuing an under-15 restriction. The European Parliament has proposed a minimum age of 16 for social media and AI companions with parental exemption opt-in for 13 to 16 year olds.

Even Ireland, which houses the European headquarters of most major social media companies, joined the wave with the government signaling last week plans for age verification measures that could effectively bar under sixteens from major platforms. *"This isn't fringe policy anymore,"* Ailís notes. *"This is [a] real accelerating shift in youth access to digital spaces."*

Alex emphasizes the pioneering nature of Australia's legislation and notes an unexpected wave of support. *"I was on TikTok last night and I saw a video of the eSafety commissioner in Australia talking about this ban,"* she shares. In the comments, she observed that most people were expressing support for the ban. *"I actually found that surprising. There was a comment to the effect of, 'as someone who belongs to Gen Z, I agree with this. This was so necessary for us, meaning Gen Z, but no one stood up for us."*

However, both experts acknowledge the enforcement challenges ahead. *"Platforms aren't going to do this perfectly,"* Alex predicts. *"They have to take reasonable efforts here. Then we also know that children are clever and they will find ways to circumvent this ban. The real question for me is knowing that children inevitably will end up online, what are platforms doing to make that experience safer?"*

Ailís points to Roblox as a critical case study. *"Unlike traditional social media, which is at least designed for 13 plus, there's no minimum age to open a Roblox account. That means that the vulnerability of their users is significantly higher and the harm some very young children are experiencing is deeply worrying,"* she explains.

But the motivation and support for change isn't always there.

"When Roblox announced it was increasing spending on safety and infrastructure, its share price dropped about 15%. Investors will treat safety, even child safety, as an overhead. I think that's quite shocking. The safest platforms should be the most valuable, but that's not how the market works."

Ailís Daly
HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

There are, however, positive developments. Alex notes that *"Roblox rolled out technology to estimate a user's age by analyzing images of their face. The goal here is to prevent children and teens from chatting with unfamiliar older users. They're going to use this data to comply with the ban in Australia."* She adds that while *"that system won't be perfect... I think it's far more effective than relying on self-reported ages and hopefully a sign of where the industry is headed."*

A critical gap in current policy concerns what happens to displaced teens. *"Losing access to where you connect with friends, communities, and resources, that can be immensely isolating, and that's the gap in current policymaking. We are removing teens without building what comes next,"* Ailís warns. *"If governments are determined to remove teens from digital spaces, there must be equal investment in offline youth programs, community spaces, and safe alternative digital environments. Otherwise, we're going to just solve one safety problem and create a loneliness crisis."*

2 The Rise of AI Companion Apps

Directly connected to social media restrictions is the predicted surge in AI companion app usage, particularly among young people.

Alex points to a tragic catalyst for increased scrutiny: *"There was that tragic case of a 14-year-old who took his own life after a character AI persona allegedly encouraged him to do so. After this incident and other lawsuits and criticism, character AI has since chosen to ban children from talking to its chatbots."*

"The prediction isn't just that AI companions will rise. It is that regulators, that parents and the public will demand a completely different safety standard for this kind of AI that imitates relationships."

Ailís Daly
HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

"We know that these chatbots have the tendency to be overly sycophantic and to make mistakes. And sadly, some of those mistakes can be deadly," Alex notes. *"I do hope here that regulation, lawsuits, and new research into the impacts of chatbots on youth will encourage AI companies to take it seriously and consider bans for certain age groups as character AI has since done."*

She stresses the importance of proactive safety measures: *"The time to invest in safety is not post-production after a tragic incident has occurred. It is actively red teaming the system. It is ensuring that you're investing in trust and safety teams that those safeguards are in place. Then of course, those efforts are ongoing. It's not a one and done exercise."*

The connection to social media bans is direct. *"Removing [teens] from their digital spaces where they've grown a presence without building anything healthier, then they will naturally turn to the tools that are in the zeitgeist and that talk back to them sometimes or oftentimes without judgment,"* Ailís observes. *"And unless we reshape those tools now, we risk swapping one set of harms for another."*

Alex highlights a fundamental problem with current companion apps: *"Historically, the experience for youth online has been largely similar to the experience of adults. They're using these apps that adults are using, they're having largely the same experience, and yet we know that they are a vulnerable population."*

"The real question is whether we've built [companion apps] to be safe for the most vulnerable users who will turn to them first."

Ailís Daly
HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

3 Humans Still Matter in Content Moderation

"Humans are still going to be absolutely essential in content moderation. AI has actually been part of moderation for almost 20 years. Early models weren't anything like today's generative AI. They were basic machine learning classifiers, but they did do and still do a lot of heavy lifting with content enforcement at scale."

Alexandra Popken
SVP OF TRUST & SAFETY AND AI SERVICES, WEBPURIFY, AN INTOUCHCX COMPANY

While acknowledging AI's growing sophistication, Alex explains the limits: *"We're entering a phase where LLMs can genuinely supercharge moderation, particularly because of their contextual, up-to-date understanding. And there's a lot of actually really cool experimentation happening here but getting there at scale, in my opinion, will take time. Policy decisions, nuance, context, intent – these aren't easy calls and a lot of moderation work is incredibly complex both on the policy side but also on the operational side. And it still is quite human dependent."*

Financial and infrastructure barriers also slow AI adoption. *"I think even if organizations wanted to go all in on AI tomorrow, most quite frankly don't have the maturity, infrastructure, or budget to deploy LLMs at true production scale. We're finding that it is quite expensive and resource intensive. That's not to say that those blockers will be the case for the foreseeable future, but I think they will continue to be roadblocks in 2026."*

Perhaps most ironically, AI moderation systems require extensive human involvement.

> *"Here's the irony – moderating AI requires humans, a lot of them. People are needed to label data sets that train the AI, to QA model outputs, to write and refine prompts, to review edge cases and red team risky scenarios. Realistically, this has actually provided a whole new category of human in the loop moderation,"* Alex explains.

Ailís' message to worried workers is reassuring: *"AI is of course transforming trust and safety, but for anyone listening who's worried about being replaced, relax, 2026 is not the year humans become obsolete. You can cancel your early retirement plans."*

4 A ‘Content Authenticity’ Crisis

"Rage bait is Oxford's word of the year. ‘AI slop’ might take the cake in 2026. AI slop refers to low quality, mass produced AI content, and it's everywhere."

Alexandra Popken
SVP OF TRUST & SAFETY AND AI SERVICES, WEBPURIFY, AN INTOUCHCX COMPANY

"Some forecasts suggest that by 2026, close to 90% of all web content could be AI generated," says Alex. *"That is staggering and, candidly, I feel that shift across platforms when I'm using social."*

The explosion of AI-generated content will force platforms to grapple with unprecedented challenges around authenticity. Beyond low-quality content, more sophisticated AI outputs pose serious risks. *"While a lot of AI content is noisy and low effort, there's also a growing wave of high quality, convincing AI output. In some cases, it's genuinely hard to tell what's real versus fake. I think that's actually a sobering thought. When you imagine the fallout of seeing, for example, a public figure appear to say something that they never said, the consequences are very real."*

From a trust and safety perspective, enforcement is extraordinarily complex.

"AI generated doesn't automatically equal harmful. Think about that pope in a puffer jacket moment from 2023. It was totally fabricated, extremely believable. But the real question is, is it dangerous? Should it be removed? Does it cause real world harm? These are the gray zones that trust and safety teams are going to be navigating for a long time – and certainly in 2026."

Alexandra Popken
SVP OF TRUST & SAFETY AND AI SERVICES, WEBPURIFY, AN INTOUCHCX COMPANY

Current solutions remain inadequate. *"TikTok, Instagram, and others have started using AI labels but the reality is that those policies and processes are still evolving. A lot of it depends upon self-reporting, and we know when we talked about age verification that self-reporting is just not an effective mechanism."*

Technical solutions are emerging, but limited. *"Google recently rolled out a Gemini feature that analyzes an image and tells you if it was created with Google AI. Now I think that's helpful, but it's not telling you whether content was created from other AI, so it's helpful only in specific use cases. I think it only scratches the surface of what's needed. Regardless, this is going to be a big focus for companies and trust and safety teams in 2026."*

Ailís emphasizes the real-world dangers beyond image manipulation: "I think about the much darker side of this and beyond even images, the really sophisticated frauds that can use AI generated voices to trick someone like my dad into transferring money and possibly using audio lifted from this very podcast. That to me is why the tech investment here needs to be enormous – and not just to curb AI slop, but to mitigate genuinely high risk harms."

Alex raises another troubling dimension: *"On the flip side of people being more trustworthy than they should be, there's also a pervasive harm if people start distrusting everything that they see online. It's like where are people then getting information? I think both sides of the equation are harmful."*

Industry standardization remains elusive. *"There's the Coalition for Content Provenance and Authenticity or C2PA that's creating standards around how we can indicate how something was created and by whom. But the problem that we tend to see in tech with regulation, with standards, is that it's adopted piecemeal,"* Alex notes. *"We see that it's enforced in different ways by different jurisdictions. We see that some companies adopt certain measures and others don't. Not only does that make it extremely difficult for the rank and file trust and safety employees who are really trying to do right by their users, but are grappling with all of these various challenges coming at them, it also makes the user experience pretty inconsistent as well."*

5 We Get AI Risk Savvy

"This is a public health issue now. We live so much of our lives online. We work, we shop, we bank, we date, we socialize, we learn, and yet our digital literacy is nowhere near where it needs to be.

So if 2026 is going to be the year of AI everything, it also has to be the year we inoculate people against AI scams, against AI manipulation. Education has to catch up with the reality that we're living in."

Ailís Daly

HEAD OF TRUST & SAFETY, EMEA,
WEBPURIFY, AN INTOUCHCX COMPANY

For Ailís, the challenge of authenticity online can't be solved by detection alone. This is partly because, in many cases, the damage is already done.

"When we talk about grappling with authenticity online, we have to be honest about where we are," she says. *"In a lot of cases, the horse has already bolted."* The tools that enable manipulation – deepfakes, nudification apps, hyper-realistic synthetic media – are no longer experimental or difficult to access. *"They're cheap, they're easy to use, and the ability to turn something benign into something harmful is now literally at everyone's fingertips."*

But even that level of technical investment won't be enough on its own. What we're dealing with now, Ailís argues, is bigger than any single platform or product decision.

"This is a public-health issue," she says. *"We live so much of our lives online. We work, we shop, we bank, we date, we socialize, we learn, and yet our digital literacy is nowhere near where it needs to be."* As AI becomes embedded across everyday interactions, the gap between what technology can do and what people are prepared to recognize becomes increasingly dangerous.

That reality raises the bar for what platforms need to do next. For Ailís, meaningful progress will require sustained, large-scale investment, not just to deal with low-quality AI content, but to prevent and mitigate genuinely high-risk harm. *"This isn't optional infrastructure,"* she says. *"It's critical safety plumbing for the internet."*

She also believes that the incentives need to change. *"I hope we reach a point where markets start rewarding companies that invest heavily in safety, rather than punishing them with a stock dip for doing the right thing. Protecting users shouldn't be treated as an overhead or a drag on growth. It's a prerequisite for long-term trust and sustainability."*

"If 2026 is going to be the year of AI everywhere, it also has to be the year we actively inoculate people against manipulation, scams, and synthetic deception. Education has to catch up with the reality we're already living in."

Ailís Daly
HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

That need for education sits alongside something more uncomfortable: a broader societal reckoning about misuse. Ailís draws parallels with other areas where clear boundaries were established not because something was impossible, but because the harm was unacceptable. *"There are echoes here of drink driving or human cloning,"* she says. *"Just because something can be done doesn't make it acceptable."*

In her view, the same norm-setting work is now required for AI. *"Creating or sharing non-consensual synthetic imagery, manipulating people through AI-generated deception, exploiting trust at scale – those behaviors shouldn't be seen as edgy or innovative. They should be socially abhorrent."*

Ultimately, Ailís frames AI savviness as a three-part challenge and a shared responsibility. The paradox is that the more powerful AI becomes, the more human oversight, literacy, and governance it demands.

> *"Grappling with authenticity in the age of AI means serious investment in safety technology, a massive uplift in digital literacy, and a cultural reset that clearly signals where the lines are,"* she says. *"Without all three, we'll keep playing catch-up."*

6 We'll Tackle Non-Consensual Intimate Imagery

"2026 could be the year we see real industry-wide progress in stopping non-consensual intimate imagery or NCII from spreading online. The technology already exists. We don't need a moonshot. We need adoption. My prediction, and honestly my hope, is that 2026 is the year protecting consent online stops being optional. It's the year that NCII prevention becomes table stakes for operating any digital platform – a year where the safest platforms aren't the exception, they're the standard."

Ailís Daly
HEAD OF TRUST & SAFETY, EMEA, WEBPURIFY, AN INTOUCHCX COMPANY

Perhaps the most hopeful prediction for 2026 is meaningful progress on preventing the spread of non-consensual intimate imagery (NCII).

Ailís highlights Stop NCII as *"one of the most promising and genuinely survivor-centered tools that is out there."* Run by *"the UK's revenge porn helpline and supported by meta and a growing network of platforms, Stop NCII protected 2 million images. Nearly double the year before, and supported hundreds of thousands of survivors globally. Regulators like Ofcom have publicly called it a leading model for proactive NCII prevention."*

The technology works through privacy-preserving hashing. *"A survivor processes the image locally, so the image never leaves their device. The tool generates hashes – these unique digital fingerprints. Those hashes are shared securely with participating platforms. If anyone tries to upload the same image, the platform can block it proactively before it's ever published. It's anonymous, it's proactive, it gives control back to the survivor."*

Current adoption includes major platforms including Instagram, Facebook, TikTok, Bumble, Discord, Reddit, Snapchat, OnlyFans, PornHub, and more. But Ailís stresses this is insufficient. *"There's actually only 17 in total. This list should be endless. If your platform has an upload image button, integrating [Stop NCII should be a default expectation because at this point, the question is no longer can we do this: we absolutely can. The real question is why haven't all platforms committed to doing this yet?"*

Additional tools complement Stop NCII. *"There's also the National Center for Missing & Exploited Children's CyberTipline, which helps remove explicit images of minors from anywhere online,"* Ailís explains. *"For young people who don't know where an intimate image may have ended up, there's 'take it down', which lets them anonymously create hash values for images of themselves, so platforms can detect and remove them even without a URL."*

AI-generated intimate imagery presents new challenges. *"We're going to see a surge in AI generated images. That will include AI generated intimate images, including deep fakes and non-consensual use of someone's image, which hash matching alone can't always detect,"* Ailís warns.

Regulatory momentum is building. *"Regulators like the UK with its 2025 AI Child Safety legislation are already requiring platforms and developers to block this material by design. In the US, new laws like the Take it Down Act are criminalizing NCII and mandating rapid removal, including synthetic DeepFakes."*

A critical barrier remains awareness. *"Reporting remains incredibly low. That stigma associated with this type of violation is enormous. Many survivors don't even know that these tools exist. I would suggest [that] if many, if not most, trust and safety professionals don't know that these tools exist, then how would survivors who don't even work in this industry know that they exist?"*

7 More Power for Users With Regulation and Dispute Resolution

"One of the biggest shifts in Trust & Safety in 2026 will be the growing empowerment of users on social media platforms due to regulation."

Anna MacCarthy Adams
SENIOR DIRECTOR OF CASE REVIEW AT THE APPEALS CENTRE EUROPE

A fundamental shift in power dynamics is coming to trust and safety in 2026, as users gain unprecedented agency over content moderation decisions.

Anna explains the historical imbalance: *"Until recently, social media users were the subject of Trust & Safety conversations, but had little say themselves. If they disagreed with a social media platform's decision to remove their account or leave up potentially harmful posts, they had few options. They could ask the platform to take another look at their decision – which often meant platforms just agreeing with themselves. The other choice was expensive, lengthy and out-of-reach for most people: take the platform to court."*

The EU's Digital Services Act has created a new avenue for redress. *"The introduction of the EU's Digital Services Act has brought a new option: out-of-court dispute settlement bodies. If people across the EU disagree with a social media platform's decision, they can ask one of these independent, certified bodies – such as Appeals Centre Europe – to review it free-of-charge,"* she notes. *"While dispute settlement bodies' decisions are not binding, platforms are required – under EU law – to engage with them in good faith, and they have already implemented many decisions."*

Early adoption signals strong demand. *"In 2025, thousands of social media users from across the EU submitted disputes to dispute settlement bodies – indicating real appetite for using this new option. In 2026, we expect to see even more people and organisations using this new right to take greater control over what they see and post online,"* Anna predicts.

She sees this creating valuable feedback loops for platforms: *"With dispute settlement bodies, like the Appeals Centre, already making thousands of decisions on social media disputes, they are amassing valuable data. As this data grows throughout 2026, platforms will be able to use it to identify and address content moderation 'blind-spots' where they are making wrong decisions. If they choose [to] do this, their users will have a better experience, benefiting the people who use their platform and, of course, the platform itself."*

8 T&S Will Shift From Compliance to Business Enablement

The conversation around trust and safety investment is shifting from moral and regulatory imperatives to demonstrable business value. AJ Brown of the Brand Safety Institute identifies this as *"less of a shift and more the continuation of a trend that we really saw pick up this year."*

"A key theme in 2026 will be the continued reframing of risk mitigation (and all of the various forms it takes) as business enablement rather than a primarily moral or compliance-driven exercise," AJ explains.

"Over the past year, we've seen growing pressure to align Trust & Safety and Media Responsibility investments with quantifiable business outcomes – revenue protection, advertiser confidence, user retention, cost reduction – and that focus will only intensify next year."

The message to trust and safety leaders is clear: *"Trust & Safety must speak the language of ROI to see continued internal support and investment in an increasingly volatile business environment."*

"A key theme in 2026 will be the continued reframing of risk mitigation (and all of the various forms it takes) as business enablement rather than a primarily moral or compliance-driven exercise."

AJ Brow
COO, THE BRAND SAFETY INSTITUTE

AJ also emphasizes the critical importance of interpretability in AI-driven systems.

As safety systems become more complex, transparency becomes essential. *"As safety systems move away from legacy blocklist-based systems and favor AI-driven enforcement, they become more complex and opaque. Given this reality, it's critical that platforms be able to clearly articulate how AI-driven decisions are made and why outcomes occur: for users, advertisers, and regulators,"* AJ notes.

"Transparency won't just build trust; it will become a competitive advantage that enables more informed, faster adaptation as the landscape inevitably shifts in ways that will make 2026 look meaningfully different from today, almost certainly in ways no one can fully anticipate."

"Given the speed at which the ecosystem is evolving, largely driven by AI, none of us should assume we can reliably predict what next year's risk landscape will look like. In this kind of environment, platforms need to prioritize explainability and interpretability alongside sophistication."

AJ Brow
COO, THE BRAND SAFETY INSTITUTE

9 The Move from Reactive Compliance to Proactive Innovation

The trust and safety field is maturing beyond merely responding to crises. Dr. Rachel Kowert of Psychgeist explains: *"While I think regulatory pressure is the underlying push for this shift, I'm happy to see companies taking the impetus to get ahead of the curve rather than scrambling to catch up."*

The gaming sector offers promising examples. *"In the gaming sector specifically, we're already seeing this manifest in some really promising ways. We're seeing sophisticated voice moderation systems being integrated at the design level rather than bolted on afterward. Companies are moving beyond text-based detection to tackle the nuanced challenges of real-time voice communication, which is where so much harassment actually happens but has historically been the hardest to moderate effectively."*

Transparency is becoming a competitive differentiator. *"We're also seeing a shift toward transparency and greater public communication about these efforts. Trust & Safety teams are increasingly engaging in public discourse about their methodologies and challenges,"* Dr. Rachel observes. *"In games, I would argue that Activision (in collaboration with CalTech) set the gold standard here by publishing research about both their successes and failures in community safety. This kind of openness not only helps the entire industry learn faster, but it also demonstrates to regulators and stakeholders that companies are taking these issues seriously as legitimate areas of research and innovation, not just necessary evils."*

"I predict a fundamental shift from only reactive compliance to investing in proactive innovation in trust and safety practices."

Dr. Rachel Kowert
FOUNDER OF PSYCHGEIST

Knowledge sharing accelerates progress industry-wide. *"When companies start sharing data about what works and what doesn't, we move from everyone reinventing the wheel to building collective expertise. That's exactly what we need to tackle these increasingly sophisticated threats."*

Dr. Rachel stresses the importance of leveraging external expertise rather than attempting to build all capabilities in-house.

"For me, the most critical adaptation platforms need to make is recognizing that they can't (and shouldn't try to) be experts at everything."

Dr. Rachel Kowert
FOUNDER OF PSYCHGEIST

She points to extremism as a prime example: *"We've been hearing a lot recently about how extremist groups are leveraging digital platforms, and this is a perfect example of why the 'figure it out internally' approach doesn't always work. Extremism research is a highly specialized field with its own methodologies, threat assessment frameworks, and intervention strategies. If you don't have that expertise in-house, bring it in. Partner with researchers, consult with organizations that specialize in prevention and intervention, engage with community groups who understand these dynamics firsthand."*

> *"The same principle applies across the spectrum of emerging risks. Whether it's understanding the intersection of gaming culture and political radicalization, navigating the complexities of AI-generated content, or addressing technology-facilitated gender-based violence - there are experts who have spent years studying these phenomena. You don't need to become world-class researchers in every domain, but you do need to be world-class at identifying and leveraging the right expertise."*

Her advice for success: *"The platforms that will succeed in 2026 are those that build agile consulting networks rather than trying to expand their internal teams to cover every possible risk area. This approach is not only more cost-effective, but it also ensures that platforms are utilizing cutting-edge insights rather than once again reinventing the wheel (which are typically less effective versions of the wheel anyway). The expertise is out there (and are ready and willing!), the platforms just have to bring them in."*

10 Users Will Actively Disengage With Toxic Platforms

"By the end of 2026, one of the most significant Trust & Safety shifts I expect to see is people actively seeking greater agency over their online environments, including choosing to log off or disengage when platforms feel too toxic."

Seyi Akiwowo
MULTI-AWARD WINNING FOUNDER AND AUTHOR OF 'HOW TO STAY SAFE ONLINE'

Beyond regulatory empowerment, users are taking control through voluntary disengagement from toxic platforms.

"This isn't about personalisation or content preferences; it's about control over exposure to hostility, intimidation, and escalation," says Seyi Akiwowo. *"As awareness grows around the impact of online toxicity on mental health, leadership, and civic participation, more people are becoming intentional about where, how, and whether they show up online at all,"* Akiwowo explains.

She urges platforms to reframe disengagement as a warning signal: *"For platforms, disengagement should be understood as an important signal that systems are failing to protect agency and dignity, not a lack of interest."*

Emerging risks include heightened threats to democratic participation. *"At the same time, emerging risks will include increased online violence directed at politicians and public figures, particularly as multiple elections take place globally. This poses serious safety concerns for individuals and wider risks to democratic participation, representation, and trust."*

Seyi advocates for a paradigm shift in platform safety approaches: *"To stay ahead of these risks, platforms will need to move beyond reactive moderation and invest in safety by design, informed by a public health approach, one that prioritises prevention, harm reduction, and population-level wellbeing rather than relying solely on enforcement after harm has occurred. This will require clearer escalation pathways, leadership-level decision-making, and a more systemic understanding of how design, amplification, and enforcement intersect during high-pressure cultural and political moments."*

11 Anticipatory Risk Management for Real-World Platforms

"The biggest Trust & Safety shift we will see in 2026 is the move from reactive, pattern-based prevention to anticipatory, real-world risk management, particularly for platforms that connect people offline."

Sarra Eddahiri
GLOBAL HEAD OF SAFETY, BOLT

For platforms that connect people in physical spaces, a new risk paradigm is emerging.

She emphasizes the unique stakes for ride-sharing platforms: *"In ride-sharing, Trust & Safety risks do not stay confined to the platform. They materialize in the real world. Traditionally, many prevention systems have relied on machine learning models trained on historical incidents to detect repeat patterns. While effective, the rapid adoption of AI tools is fundamentally reshaping the threat landscape."*

AI accelerates the evolution of threats. *"Malicious actors can now use AI to test, adapt, and scale harmful behaviors far more quickly than before. New abuse patterns, ranging from fraud to harassment to coordinated misuse, can emerge and spread globally before platforms have sufficient historical data to respond. This is especially critical for ride-sharing platforms, where even a single failure can result in serious physical harm."*

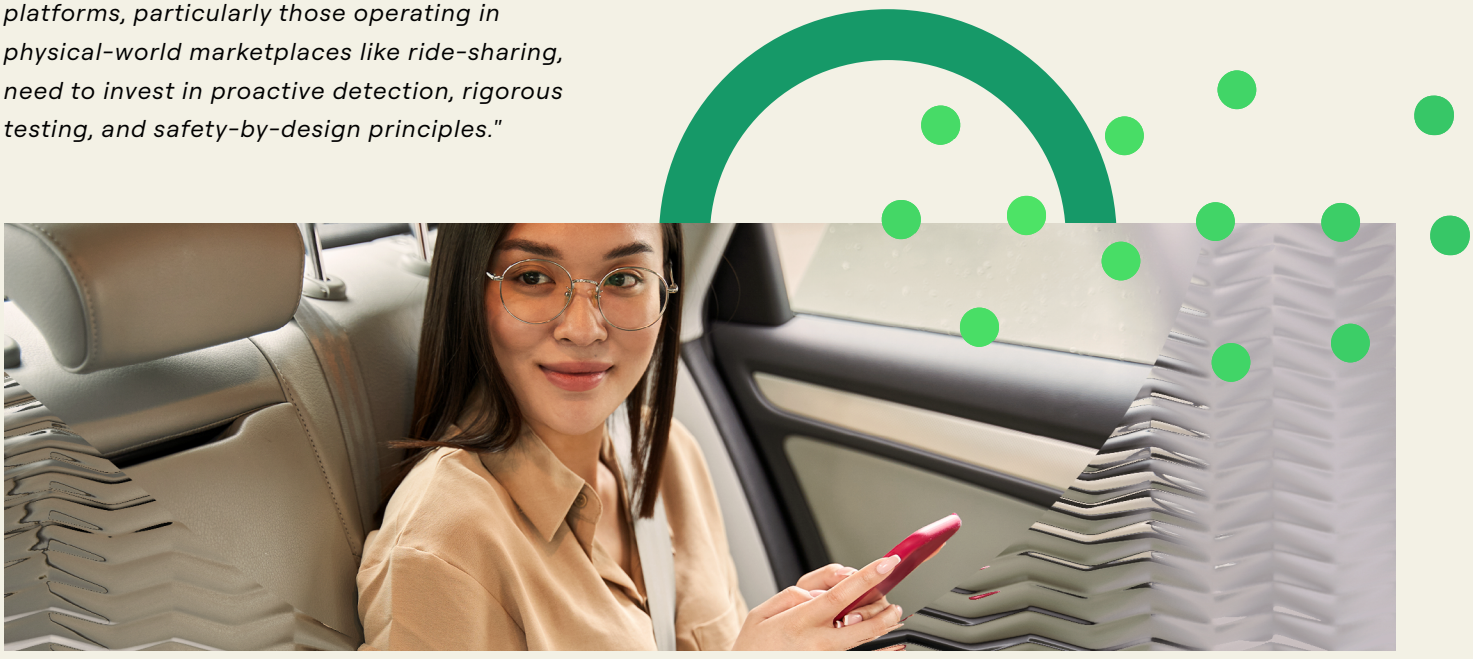
The solution requires forward-looking detection: *"This shift is driven by the need to move beyond learning only from past incidents and toward identifying early signals, behavioral anomalies, and product-level vulnerabilities before they escalate into real-world harm."*

Sarra outlines concrete adaptation strategies: *"To stay ahead of emerging risks in 2026, platforms, particularly those operating in physical-world marketplaces like ride-sharing, need to invest in proactive detection, rigorous testing, and safety-by-design principles."*

> *"This begins with stronger data science capabilities and more advanced detection models that can quickly identify abnormal behavior patterns that signal emerging risks, not only confirmed policy violations. At Bolt, this means continuously refining how we detect unsafe patterns throughout the entire ride lifecycle, from account creation to post-tripactivity. Advanced safety features such as trip anomaly detection and the identification of unsafe areas provided tangible benefits and have earned strong recognition from customers and drivers."*

Red teaming is essential. *"Actively stress-testing products and features helps identify how tools that connect people in real life could be misused, allowing teams to address vulnerabilities before launch or scale."*

Safety must be foundational, not supplemental. *"Finally, safety must be embedded into product design from the outset. For ride-sharing platforms, this requires assessing Trust & Safety risks early and ensuring that mitigation measures such as verification, controls, and safeguards are built into new features by default, rather than added after harm has occurred."*



12 A Watershed Moment for Violence Against Women and Girls

"I predict a new focus on Violence Against Women and Girls [VAWG] that is surpassing the previous frictions that have stalled progress in combatting GBV [gender-based violence]."

Emily Harman
TRUST & SAFETY EXPERT AND NCII ADVOCATE

Emily Harman identifies a significant regulatory and cultural shift that centers on technology-facilitated violence against women.

She points to landmark US legislation as evidence of changing priorities: *"The TAKE IT DOWN Act coming out of the US, a country with a huge civil liberties and privacy preserving culture, showcases that. Other rights and sensitivities are being sidelined (quite correctly in my opinion) in acknowledgement that more needs to be done to eradicate the threat to women from tech facilitated violence."*

The UK is taking a multi-pronged approach. *"The UK'S VAWG strategy has also called for early interventions within schooling to curtail the development of misogyny. Alongside this we've seen the ratification of the UN Cybercrime Treaty which mandates the criminalisation of each member state of intimate image abuse - a form of sexual assault which disproportionately affects women."*

Multiple regulatory initiatives will converge in 2026.

Emily's guidance for platform adaptation is unequivocal: *"Future proof policies to allow for global harmonisation, design regulation-agnostic risk assessments, lean into partnerships with law enforcement, civil society and academics to deepen intelligence around risk to your user base, implement recommendations even if they aren't mandated [by] law and shun a minimum viable compliance approach - instead favouring a maximised safety infrastructure."*

"[In] 2026 we'll see the enforcement of the TAKE IT DOWN act, the implementation of the UK'S VAWG strategy and a trickle down effect of the Cybercrime Treaty with more countries criminalising NCII offences. 2026 will also see the introduction of Ofcom's Additional Safety Measures. Among other things this will likely lead to mandatory hashing for NCII across platforms within certain categories - leading to an estimated 10k platforms affected by this mandatory measure."

Emily Harman
TRUST & SAFETY EXPERT AND NCII ADVOCATE

13 T&S Becomes A Systems-Level Leadership Function

In 2026, Trust & Safety will no longer be judged primarily by how much content it removes, but by how effectively it prevents harm from occurring in the first place.

For Sophie Walsh, Director of Trust & Safety at Depop, the shift ahead is not simply better tooling or faster enforcement. It is a fundamental change in how Trust & Safety functions are positioned inside organizations, moving from downstream moderation teams to strategic leaders shaping safer systems at scale.

"Trust & Safety teams will be expected to anticipate harm before it materializes, rather than primarily responding after violations occur," Walsh explains. *"That means embedding safety considerations earlier in product design, model development, and policy decisions, and measuring success by harm reduction, not just enforcement volume."*

And several forces are converging to make this shift unavoidable.

"The biggest Trust & Safety shift in 2026 will be the move from reactive enforcement to proactive, systems-level risk management."

Sophie Walsh
DIRECTOR OF TRUST & SAFETY, DEPOP

Perhaps the strongest of all is the rise of generative AI, which is accelerating the scale and speed of harmful behavior, as well as its ambiguity. As AI technology increases both the volume and sophistication of abuse, the old ways of reactive moderation alone become structurally insufficient. Those platforms that rely primarily on post-hoc enforcement will find themselves permanently behind the curve and constantly responding to yesterday's threats while new ones emerge.

What's more, regulation is reinforcing this reality. More countries are putting measures in place, and frameworks such as the EU's Digital Services Act are pushing platforms toward demonstrable risk assessments, mitigation planning, and governance maturity. The expectation is no longer just that platforms respond to harm, but that they can show the foresight to identify potential risks and address them before rollout.

Cost and sustainability pressures are also driving this change. Large-scale human moderation is neither economically nor psychologically sustainable, particularly as content volumes continue to rise in the wake of generative AI. Leadership teams are increasingly demanding prevention and clearer prioritization, not just more reviewers.

But at the same time, public trust expectations are rising. Around elections, children's safety, and crisis events, stakeholders now expect platforms to demonstrate foresight, not just clean-up.

When harm occurs, the question is no longer only *what did you take down?* but *why was this possible in the first place?*

"Trust & Safety becomes a leadership and design discipline, not a downstream enforcement function. The platforms that lead in 2026 will be the ones that build safer systems by default, not the ones that remove the most content after harm has already happened."

Sophie Walsh
DIRECTOR OF TRUST & SAFETY, DEPOP

For Trust & Safety leaders, this reframes the role entirely.

In 2026, Walsh predicts, Trust & Safety will become more strategic and deeply embedded across organizations, working in close alignment with product, engineering, legal, and communications teams. Clear decision rights, escalation mechanisms, and governance structures will matter just as much as policy language or detection accuracy.

To stay ahead of emerging risks, platforms will need to institutionalize forward-looking risk assessment, moving beyond static policies toward continuous risk sensing. This includes horizon scanning for new misuse patterns, pre-launch risk reviews for features and models, and explicit ownership for emerging threat domains such as AI-enabled fraud and social engineering.

Safety-by-default design will be critical. System-level safeguards – friction, rate limits, provenance signals, and default protections for high-risk users – reduce harm before moderation is even required. When enforcement is needed, it must be adaptive rather than rigid, pairing AI tooling with rapid iteration loops, human escalation paths, and strong quality measurement.

Just as importantly, Walsh emphasizes governance and credibility. Effective Trust & Safety in 2026 will depend on tight cross-functional coordination and external intelligence. Partnerships with peers, researchers, and civil society will play a growing role in early signal detection and legitimacy, while transparency and evidence-based reporting increasingly differentiate credible platforms from reactive ones.

Taken together, this marks a defining evolution for the field. In 2026, Trust & Safety becomes about shaping the systems, incentives, and leadership decisions that determine whether harm occurs at all.