Teleperformance
trust & safety

T&S POLICY INSIGHTS NEWSLETTER

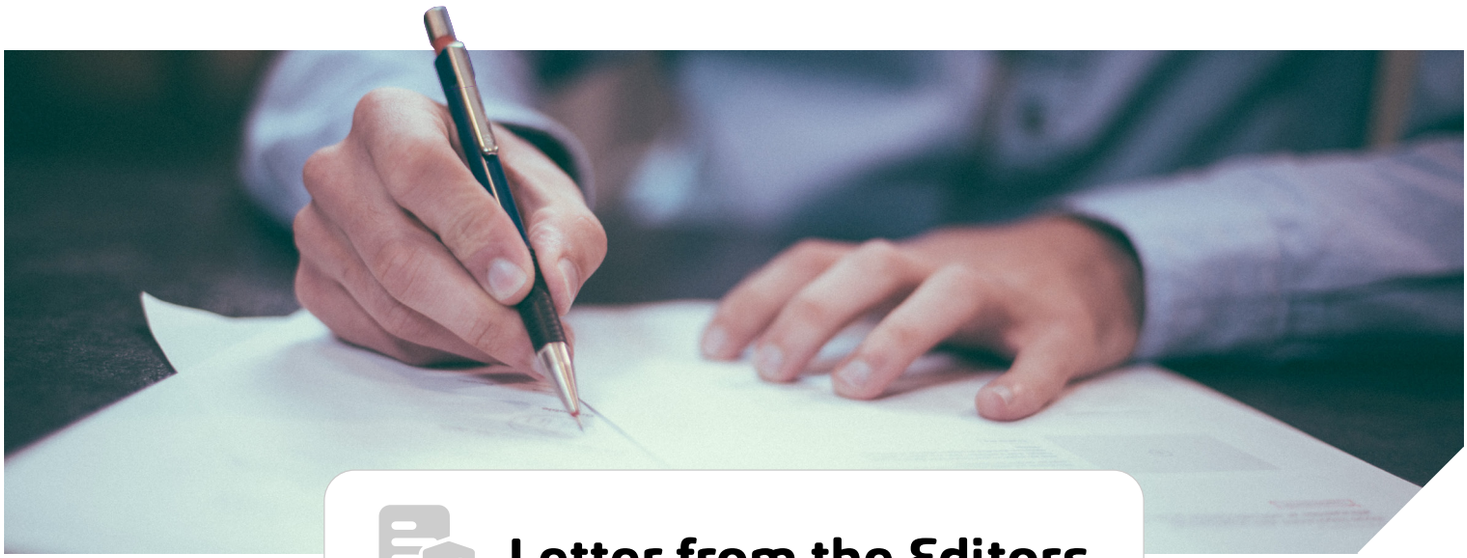
Inaugural Issue: February / March 2023

In this issue:

We highlight some of the most important policy stories in the world of Trust & Safety, including news from Europe, UK, and Australia. We also highlight new technologies and tools that could have wide ranging impacts on platform policies and moderation capabilities.

For more information contact:

Akash Pugalia, Global President, Trust & Safety and Farah Lalani, Global VP, Trust & Safety Policy.



Letter from the Editors

The policies, or "community guidelines", by which platforms operate are changing rapidly, both in their coverage and in their enforcement in order to keep up with changing cultural norms, global political unrest, new bad actor behaviors, risks posed by new technologies, and a wave of new regulation. Here are the trends we are seeing when it comes to platform policies and content governance:

1 Policies and Enforcement Strategies Will Need to Adapt to Address Deep Fakes and Other Harmful Content Produced by Generative AI

Generative AI technologies have risen in popularity, and with it, concerns about how effective policies to deal with harmful content spread through AI will be created and enforced. Given that some experts predict as much as 90 percent of online content could be synthetically generated within a few years, this is no small problem. So, what are some the challenges that Generative AI poses?

- One main challenge is that it is difficult to distinguish AI vs. human-generated content in the first instance. For example, some research has shown that people cannot reliably discern pictures of real faces from computer-generated images. This could have several negative implications, for example, the risk that this could lead to the proliferation of more realistic deepfakes.
- Generative AI models that produce images, pose both privacy and copyright challenges. These technologies can be prompted to produce identifiable photos of real people, unwanted "nudes", and/or other content that people feel violate their rights. The legal and ethical dilemma this poses for intellectual property rights is unclear given the way that these models use content without citing or attributing original sources.
- While output from Generative AI models can sound very convincing, the content is not always accurate or true. The way these models present information can make it seem authoritative, which make it a potential vector for harm, especially if it is being used by bad actors to manipulate public perception, through astroturfing campaigns.

The likely risks from generative AI will evolve as adoption grows and the technology is integrated in mainstream digital products and services. Starting points for mitigating these risks could include better understanding the sources of content used in these models, developing governance models around its use, and creating avenues for transparency when such models are used. There is a long road ahead to managing the substantial risks of generative AI technologies and ensuring that policies account for these enhanced risks.

2 With New Regulation, Increased Transparency around Platform Policies has become 'Table Stakes'

The new regulation emerging across the globe – while different in their scope, implementation, and underlying principles – seem to have a common thread focused on raising meaningful transparency and accountability of platform policies and their enforcement.

- With Europe's Digital Services Act (DSA), platforms will need to provide users with explanations ('statement of reason') indicating why their content was actioned, what policy was violated, whether the content was deemed illegal, and how the content was discovered. This is just one example of how the DSA mandates increased transparency; more about how the DSA impacts content moderation operations is included in this issue.
- In Australia, platforms are required to report on how they are implementing the Basic Online Safety Expectations (BOSE) under the Online Safety Act 2021. This newsletter highlights a recently published article with top insights from the first-ever industry transparency report issued by the eSafety Commissioner.
- The draft Online Safety Bill in the UK now includes criminal liability for tech executives who do not comply with their duties to protect children online; proponents believe this gives the UK regulator, Ofcom, more teeth when it comes to enforcement.
- China has regulated the use of Generative AI technologies through a number of new rules; the use of watermarks is one interesting tool the government has implemented.
- In India, three Grievance Appellate Committees (GAC) were formed to address user grievances against social media companies; with the platform to submit grievances expected to be operational beginning March 1, 2023, the government has outlined how this process will better provide internet users with redress with issues they face online.

3 Cultural norms, political events, and social movements are driving new conversation around platform policies

Are current platform policies related to nudity inclusive enough for all groups of people, particularly intersex, non-binary and transgender people? There has been increased conversation around whether platform policies around nudity need to be updated, particularly in allowing depiction of bare chests in the context of health-related discussions.

In other platform policy news, the ability for users to call for violence against political leaders is being scrutinized, particularly where such calls are "rhetorical slogans" and not "credible threats".

We hope you find these insights and the round-up of T&S Policy News below to be useful!

Signed:

FARAH LALANI
Global VP, Trust & Safety Policy

AKASH PUGALIA
Global President, Trust & Safety

Platform Policies & Content Governance

Safer Internet Day 2023 – Bolstering the Fight to Protect Children Online

A deep dive into the first-ever industry transparency report released by the eSafety Commissioner showed a highly varies set of policies and practices when it comes to tackling Child Sexual Abuse Material (CSAM); this was particularly the case with using technology to detect new CSAM and CSAM on livestreams and video content. In this article in the World Economic Forum Agenda, Akash Pugalia, President of Trust & Safety and Farah Lalani, VP Trust & Safety, Policy, at Teleperformance provide a perspective on some of the major gaps in protecting children online and some of the solutions to move forward.

A Deep Dive into Content Moderation and its Role in Addressing Harmful Content Online

Recent thought leadership produced by MIT Technology Review Insights, "Humans at the center of effective digital defense", in association with Teleperformance, highlighted how trust & safety policies, practices, and technologies are evolving. Providing insight on the future of content governance, Julie Owono, Executive Director of Internet Sans Frontières (Internet Without Borders) highlighted her views in this white paper: "Content moderation rules and practices, which I include under the umbrella of content governance, will evolve. Under increasing pressure from users, advertisers, and governments for safe online spaces, we may see the emergence of more common standards and, perhaps, common procedures. This will require a multistakeholder approach through which industry, civil society, governments, and academia collaborate." Read and download the full paper here.

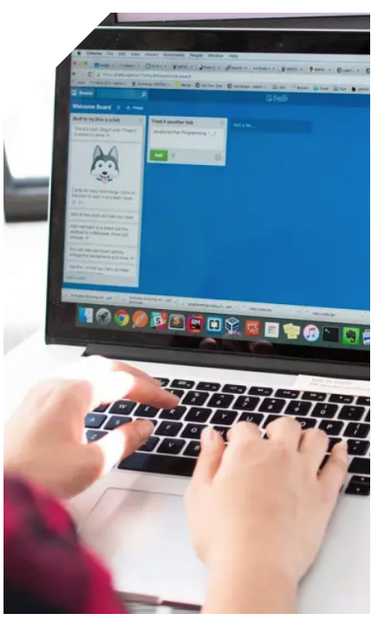
Regulatory Highlights

Section 230 in the Hands of the Supreme Court

Does amplifying an illegal post make companies responsible for it? With the Gonzalez v. Google case in the hands of the Supreme Court, the question of liability for algorithmic recommendations is at stake. Other key policy decisions in the U.S. are outlined here.

UK Strips "Legal but Harmful" from Online Safety Bill

The UK has stripped the provisions related to material that is "legal but harmful" for adults. Mark MacCarthy's adept analysis in Brookings highlights the pros and cons of the changes, which will now require social media companies to remove content only if it is illegal or violates their publicly announced standards, to have in place systems to enforce their publicly announced standards, and to provide an appeals process for users whose content has been removed.



How the Digital Services Act (DSA) changes Content Moderation

With sweeping new legislation that aims to curb the spread of illegal content now passed in Europe, questions about how this practically impacts content moderation processes, policy enforcement, and transparency are top of mind. This article written by Akash Pugalia, President of Trust & Safety, and Farah Lalani, Global VP Trust & Safety Policy at Teleperformance highlights some of the top considerations.

Policy Insights

New Policies and Enforcement Strategies may be Required for Generative AI

How should disinformation and other harmful content created by Generative AI technologies be moderated? Will it even be possible to distinguish content generated by AI from content written by people? This article takes a deeper look at the challenges that lie ahead for new policy creation and enforcement amidst these disruptive new technologies.

Reducing Polarization Online

New research has revealed how social media platforms can surface more positive interparty contact, prioritize content that's popular among disparate user groups, and take other steps to reduce polarization on social media. This is important research, showing that maybe things are not as polarized as we perceive them to be. A note of optimism for 2023!

"You can have quality, speed, or transparency with content moderation, but you can rarely have it all at once"

So you think you can moderate content?

by Katie Harbath

Policy Tools & Resources

Tech Policy Atlas

A new, very cool interactive map was released by the Australian National University and Tech Policy Design Centre. The Global Tech Policy Atlas is a very useful public repository of national tech policy, strategy, legislation and regulation, covering a range of topics including competition, online harms, digital economy, and consumer protection across a wide range of countries. At a time with a multitude of fast moving regulations, this is a handy resource for all.

New Content Moderation Tools being Developed and Shared Across the Industry

A number of tech platforms have recently made tools to tackle online terrorist content freely available to the industry. In light of the growing sophistication of bad actors, this type of cross-industry collaboration is going to be crucial to raising the safety baseline across the digital ecosystem.


Teleperformance
trust & safety