NISOS

Webinar Brief

How Security Teams and Counsel Can Successfully Navigate the Complex Challenges of Security, Trust, and Safety

Executive Summary

Over the last 15 years, technology platforms and marketplaces defined "Trust and Safety" as the protection of consumers using cross functional tactics to address gaps between physical, cyber, third-party, and fraud investigative techniques.

These methodologies serve to prevent, disrupt, and respond to any form of harm, including abuse of an online consumer, danger to company employees, and negative impact on brand reputation. All companies doing business on the internet should use similar strategies to protect their consumers or clients.

Many of the same techniques that protect confidentiality, maintain integrity, and defend data, systems, and networks (traditional cyber threat intelligence) apply to protecting consumers and are applicable to all companies, not just technology companies.

Trust and Safety PlayBook 101:

Trust and Safety teams generally consist of content moderators who are responsible for policing the platform to confirm violations exist prior to initiating takedowns.

To balance the reduction in harmful content with the need to maximize user freedoms to transact in accordance with the platform's rules, Trust and Safety teams must reflect on the mission of the brand and overall objective to bring people together and address a need through digital communication and collaboration.

Trust and Safety teams take a multi-source approach to combating threats. The following sources of data are used to determine appropriate action:

1. **User reports:** reports consumers submit about problems and threats to/on the platform.
2. **Social media:** research and automation of threats active on social media platforms.
3. **Internal tools and algorithms:** internal alerts and models of proactive defensive behavior.
4. **External experts:** law enforcement, non-profit organizations, and external intelligence vendors that can identify malicious behavior.
5. **Collaboration with other companies:** feedback and insights from other similar companies that face the same threats to brand, employees, and consumers.

Evaluating User Generated Content, Creating Harm, and Legality Concerns:

User-generated content surfaces new threats to consumer interaction with platforms on the internet. The following are areas of concern for Trust and Safety teams:

- Theft of intellectual property
- Fraud
- Harassment and bullying
- Access and equality
- Disinformation
- Physical violence

Non-traditional cybersecurity concerns may include defense against:

- Bots used to scrape data
- Service disruptions (DoS attacks)
- Spam and phishing
- Counterfeit applications
- Viruses (malware or scripts for command and control)

Counterfeit and fraud are nuanced. An actor selling fake commodities on an e-commerce platform is one type of threat that, depending on the scale, probably plays out in a game of account takedown whack-a-mole. Whereas a counterfeit application that is hijacking legitimate customers and presenting a risk to the brand's business requires more extensive legal considerations.

Section 23O:

Section 230 is the United States Communications Decency Act that generally provides immunity for website platforms hosting third-party content. The law was originally meant to protect companies from millions of users uploading content every minute of every day.

Since companies cannot edit user-submitted content, they are now being more proactive in addressing malicious threats by adjusting terms of service and more aggressively moderating content creators.

About Nisos

Nisos is the Managed Intelligence™ company. Our services enable security, intelligence, and trust & safety teams to leverage a world-class intelligence capability tailored to their needs. We fuse robust data collection with a deep understanding of the adversarial mindset delivering smarter defense and more effective response against advanced cyber attacks, disinformation and abuse of digital platforms. For more information visit: **www.nisos.com**