conectys

WHITE PAPER

TRUST AND SAFETY

Ultimate Services Review 2024

Modern Trust and Safety strategies: addressing cyber threats, combating user-generated content abuse, and ensuring legal compliance

Table of Content

Introduction

In today's digital era, with numerous cyber businesses emerging and others significantly expanding their online presence on an unprecedented scale, the complexity and diversity of threats to virtual spaces' security, integrity, and reputation are rising. This poses new challenges for digital services, necessitating a proactive and adaptive approach to safeguarding user experiences, shared data, and virtual assets.

Additionally, with the rapidly growing prevalence of user-generated content, more vulnerabilities may affect platform visitors' peace of mind, necessitating heightened vigilance and unconventional protective actions. The final touch is the imperative to confront the ever-expanding regulatory landscape, where compliance is becoming more demanding and strategic than ever.

In light of this, outstanding Trust and Safety is paramount to maintaining a responsible and trustworthy environment where users feel safe, enjoyable, and free of abuse, fraud, and harassment, expressing a desire to return for more positive moments. It is also a cornerstone to running responsible, ethical, and lawful online operations, fostering loyalty and retention, leading to tremendous popularity and significant revenue growth.

However, building an impactful and effective T&S strategy requires a nuanced understanding of the evolving digital trends, embracing novel methods, leveraging innovative technologies, integrating unconventional safety measures and moderation techniques, and providing skilled and knowledgeable talents.

How to address these challenges?

- This white paper offers suggestions and recommendations, exploring the state of Trust and Safety, analysing its evolution, and highlighting critical challenges digital platforms face.
- It also provides insights into building and maintaining an effective, customised Trust and Safety framework.
- Finally, the document looks ahead to the future of Trust and Safety and its implications for online services, particularly considering the advancing AI revolution.

All the mentioned services are highly effective and important components of the overall Trust and Safety strategy, which aims to protect online users from threats that may come in many forms, such as cybercrime, malicious activity, harmful behaviour, and more. Yet, it is a partial T&S set, and its final configuration ultimately encompasses many more elements, such as various policies, critical processes, cutting-edge technologies, and diverse tools that work together as a safety measure environment, maintaining online security and user comfort.

Nevertheless, the initiative has significantly evolved over the last decade. The Trust and Safety strategies that have worked well so far needed ongoing improvement in the face of a rapidly changing landscape, fast technological advancements, online user-generated content (UGC) growing popularity, heightened accessibility and interest in virtual activities, the rise of sophisticated cyber threats, and various challenges related to digital safety and privacy. These factors made businesses reshape Trust and Safety, which became more comprehensive and provided a higher level of protection than just a few years ago.

Below is a brief comparison illustrating a characteristic of typical T&S components from the past and today, and understanding these differences can help in decision-making on building the impactful strategy of the future:

Trust and Safety a Decade Ago	Trust and Safety Nowadays
Services were primarily focused on implementing basic security measures to protect against common online threats, like data breaches, password vulnerabilities, network intrusions, or simple malicious software or attacks.	T&S is typically a well-crafted and managed strategy prepared to address a large amount of comprehensive online risks and challenges reactively and proactively on a global scale.
Regarding content moderation, it relied mainly on manual reviewing and reactivity when individual administrators managed moderation on web pages and chat rooms.	The initiative encompasses broader services and cutting-edge technological solutions, including AI-driven tools for better threat detection and prevention.

Trust and Safety Nowadays

With the rise of social media, the need for dedicated teams to moderate user-generated content has surged. Content moderation, now a crucial Trust and Safety element, spans various formats such as text, images, videos, and voice content. Digital businesses strategically invest in assets like content policy teams, moderation teams, language capabilities, and automation technologies to provide large-scale, efficient services, meeting customer demands for online safety amidst evolving regulations..

Law compliance in the domain of Trust and Safety has transitioned to become more global and very specific in terms of the risks, obligations, penalties or groups of individuals requiring particular protection, like children and teenagers.

Moreover, it is worth mentioning that the ongoing T&S evolution has been additionally inspired by growing user demand for enhanced and risk-free online experiences, placing significant responsibilities on companies to establish a more robust and adaptive framework to ensure their online platforms' and visitors' safety. One critical aspect, for example, is the cautiousness surrounding data sharing. According to a McKinsey survey of 1,000 North Americans, 82% of consumers would avoid a company with security concerns, and 72% would cease doing business if their sensitive data was shared without permission.

Why is digital space so vulnerable to cybercrime and violations?

While being a cross-border space where people increasingly interact with each other, share content, access various groups and materials, and engage in many different activities like shopping, finance management or learning, the digital world is not immune to multiple types of threats.

Much like crimes in the physical universe, cybercrimes and violations can manifest in the virtual realm, including, for instance, unethical or aggressive activities, hate speech, harassment, cheating, and other forms of misconduct, as well as unfair practices, terroristic threats, or diverse organised illicit activities financially motivated, exploiting the vulnerabilities inherent in digital space. Their consequences can extend beyond that sphere, causing emotional, health, moral, financial, or reputational harm to individuals, communities, organisations or even nations.

This happens because a digital realm attracts individuals with varying levels of honesty, ethical values, intentions, and approaches to social norms and legal regulations. Additionally, users can often occur under pseudonyms, spanning diverse locations and jurisdictions, making it notably difficult to identify and track virtual personas suspected of rule violations or criminal actions, especially when notable groups are involved, supported by genuine experts in the intricacies of digital anonymity and manipulation.

What also encourages cyber criminals is that they can exploit software, hardware, or network infrastructure vulnerabilities, taking advantage of insecurity or advancements to carry out attacks strategically. In such cases, while technology helps businesses protect themselves, it empowers hackers to adapt, evolve and enhance the complexity of illegal tactics, manipulating security gaps and targeting individuals. Additionally, when offensive information and materials are not removed from the digital space, people can feel motivated to copy such behaviours, as it is tolerated without repercussion.

The scale and regular growth of cybercrime and violations are well reflected in many well-illustrated statistics, like, for instance:

Cybersecurity Ventures projected global cybercrime damages to reach 8 trillion U.S. dollars in 2023, a 170% increase from 2015. If considered an economy, it would rank as the world’s third-largest one, surpassing all nations except the U.S. and China.

According to the 2023 Anti-Defamation League (ADL) survey, 52% of respondents reported experiencing online harassment or hate, indicating a 12% increase compared to the previous year.

In 2021, Cybersecurity Ventures predicted that businesses faced a ransomware attack about every 11 seconds, up from 14 seconds in 2019. The frequency of attacks is expected to increase to approximately every two seconds by 2031.

In 2022, Microsoft observed 1,287 password attacks every second, totalling more than 111 million attacks daily and 3.4 billion annually, 26% more than in 2018.

Addressing all these vulnerabilities requires a comprehensive Trust and Safety approach. This involves allocating appropriate talents, skills, and a technological stack and gaining and developing expertise in critical areas such as cybersecurity, data protection, content moderation, and regulatory compliance. It also entails establishing relevant processes, defining clear protocols, and rigorous enforcement of robust legal frameworks to detect and prosecute cybercriminals successfully. Engaging in broader moderating efforts, educating users, and exhibiting agile adaptation to unforeseen situations are crucial aspects of the whole undertaking.

With outstanding T&S strategies, organisations can uphold their positions and efficiently deal with fundamental risk management considerations of the digital landscape, where different types of threats prevail. In the worst-case scenario, the absence or delivery of poor Trust and Safety measures can lead to severe consequences, including brand harm, diminished user retention, legal repercussions, or business failure.

Key challenges in building a modern Trust and Safety framework

Building a modern and impactful Trust and Safety framework from the beginning poses multifaceted challenges.

It requires a blend of human skills to handle inappropriate content effectively. Simultaneously, it involves implementing and managing robust technological solutions and support resources to swiftly prevent, identify, and mitigate various online risks. Balancing privacy, security, and freedom of speech has become a significant dilemma for digital businesses, evoking varied emotions from stakeholders on both ends of the spectrum. Some prioritise safety, while others advocate a more liberal approach, emphasising individual rights over manageable protection.

In this intricate and demanding environment, appropriately navigating these issues demands relevant strategies. Among the current main concerns to be considered and addressed are:

1. TECHNOLOGICAL ADVANCEMENTS

Technological advancements pose challenges for digital entities, necessitating continuous adaptation in Trust and Safety's efficiency, productivity, and protection levels to maintain robust and secure operations while keeping pace with the dynamic nature of online risks and staying in compliance with evolving regulations. However, the key is constant trend monitoring and a deep understanding of the potential of new solutions to ensure that the advancement makes sense and aligns with the overall strategic objectives.

The technological innovations that should be of great interest in T&S include artificial intelligence, cybersecurity technology, data privacy measures, automated content moderation tools, image and video recognition algorithms, blockchain technology, user authentication solutions, online risk monitoring programs and regulatory compliance software.

Generative AI: Particularly noteworthy is the concept of generative AI, designed to enable machines to create content, such as text, images, music, audio, and videos, in a human-like way. Such models use uploaded datasets to identify patterns and produce new and original materials based on the learned information. But, in the context of Trust and Safety, generative AI can also create potential risks in areas such as the manipulation of information, privacy breaches, and security threats. For instance:

Generative AI can produce incorrect content, a phenomenon known as AI hallucination. This is because, despite its exciting opportunities, it cannot fully imitate humans in nuances, language intricacies, and cultural understanding.

Generative AI can efficiently create highly realistic and deceptive content, including deep fakes. This technology, if not recognised and eliminated, can contribute to the spread of misinformation, manipulation, and even harm to an individual's reputation through harassment.

A significant cybercrime concern with generative AI is the risk of oversharing personal information or exposing sensitive data during training, potentially exploited by attackers for malicious purposes. Unfortunately, effectively training AI to mitigate these threats efficiently remains a challenge.

When deciding to embark on the AI-driven journey, it is crucial to create the best practices for secure use and unexpected vulnerabilities, allocate high-quality data, and prepare infrastructure, resources, and procedures such as process monitoring, safety audits, or overall management.

2. User-generated content popularity and its impact on the rise of inappropriate behaviour online

As per Grand View Research, “The worldwide market for user-generated content platforms reached a value of USD 4.4 billion in 2022, and it is expected to grow at a compound annual growth rate (CAGR) of 29.4% from 2023 to 2030”. It all means that more UGC content can be anticipated, spanning various formats, such as text, image, video, podcast, interactive multimedia, etc.

Diverse contributors, including loyal users, brand influencers, employees, field enthusiasts, and various other UGC creators, play unique roles in the UGC evolution, enriching the digital landscape and fostering interest. On the one hand, it provides significant opportunities for web services and brands, such as increased popularity and greater brand familiarity. On the other hand, more severe threats are likely to emerge, given that user-generated materials are typically published without complete control, potentially leading to inappropriate information in some cases.

Mitigating this risk is possible with continuous, strategic, and highly professional supervision through the next generation of content moderation. This, in turn, requires constant adaptation, for instance, by leveraging machine learning algorithms for greater efficiency, scalability and analysis while also dedicating more professional human moderators to the nuanced and delicate content curation. Lastly, it is also critical to consider that handling UGC cannot be a one-size-fits-all solution - it must be contextual and appropriately adapted to the digital space specificity, regulations, user groups, preferences, demographics, languages spoken or cultural backgrounds.

3. The rise of sophisticated cyber threats

As per Grand View Research, “The worldwide market for user-generated content platforms reached a value of USD 4.4 billion in 2022, and it is expected to grow at a compound annual growth rate (CAGR) of 29.4% from 2023 to 2030”. It all means that more UGC content can be anticipated, spanning various formats, such as text, image, video, podcast, interactive multimedia, etc.

In the past, when online engagement options and web surfing were more limited, security incidents occurred less frequently, their impact was comparatively smaller, human-empowered attacks were prevalent and more straightforward, and the scope of potential harm was narrower. In contrast, while so many aspects of our lives have gone digital, including business, education, daily routine or entertainment, cyber threats nowadays are more persistent, complicated, of a global nature, and often orchestrated by complex entities and increasingly driven by AI capabilities, contributing to more severe damages and consequences.

Below is a comparison outlining key characteristics of potential digital threats over time, illustrating the rise of more sophisticated issues impacting virtual reality today:

Potential Threats a Decade Ago

- **Data breaches:** Unauthorised access leading to the exposure of sensitive data
- **Hacking:** Gaining unauthorised entry to computer systems or networks
- **Privacy violations:** Breaches of individuals' privacy through unauthorised actions.
- **Malware:** Infecting computer systems with malicious software, including viruses, worms, and trojans.
- **Ransomware:** Blocking access to personal data unless a ransom is paid
- **Phishing attacks:** Phishing techniques used to trick individuals into revealing sensitive information.
- **Denial-of-Service (DoS) attacks:** Intentional actions to disrupt computer resources or networks, making them inaccessible to users.
- **Social engineering attacks:** Manipulating individuals to share confidential information.

Today, additional challenges encompass expanded cyber espionage, sophisticated phishing attacks, and the impact of emerging technologies such as quantum computing on recent encryption methods.

Potential Threats Nowadays

- **Data breaches:** Sophisticated and widespread breaches posing heightened risks to sensitive information's confidentiality and security
- **Hacking:** Gaining unauthorised access to computer systems, often on a larger and more sophisticated scale, targeting critical infrastructure and networks
- **Privacy violations:** More targeted and invasive privacy breaches, usually initiated by organised groups
- **Cyberattacks on IoT devices:** Targeting vulnerabilities in internet of Things devices
- **Targeted attacks:** Focused and personalised cyber threats against specific individuals or entities
- **Supply chain attacks:** Exploiting vulnerabilities in the supply chain to compromise systems
- **Ransomware-as-a-Service:** Offering the SaaS services to other cyber criminals who want to carry out ransomware attacks
- **Biometric data theft:** Unauthorised access or theft of biometric information
- **Cryptocurrency theft:** Illicit acquisition of digital currency
- **Spyware:** Using malicious software to spy on a user's activities
- **Formjacking:** Injecting malicious code into online forms to steal sensitive information
- **Pharming:** Redirecting website traffic to fraudulent sites for malicious purposes
- **Zero-Day exploits:** Vulnerabilities in software or hardware that attackers exploit before the vendor releases a patch or solution

4. Legal consideration

One of the paramount challenges in Trust and Safety revolves around legal and ethical considerations, which are constantly and significantly evolving. They increasingly focus on enhanced digital privacy and protection, fostering transparency and adapting to changing technologies while holding online platforms accountable for ensuring a safe environment free from virtual abuse. For instance, in recent years, numerous new regulations have been introduced across regions, shaping the digital landscape and emphasising the importance of online safety. Key legislative initiatives globally include:

United Kingdom

The UK Online Safety Bill is a comprehensive regulatory framework that holds social media platforms accountable for the content they host. It obliges them to promptly address online harms, enhance transparency, and ensure a safer digital environment for users in the United Kingdom, with a particular focus on children’s protection.

European Union

The documents such as the Digital Services Act and the Code of Practice on Disinformation represent the European Union’s commitment to fostering a secure, fairer and more transparent online space, free of misleading content and fake news.

United States

The US Kids Online Safety Act (KOSA) and California’s Age-Appropriate Design Code Act (CAADCA) aim to enhance online safety for children in the United States, focusing on age-appropriate content and design standards.

Singapore

Introduced by the Singapore Parliament, the Online Safety Bill reflects the global trend of countries enacting legislation to address online safety challenges and protect users.

India

Indias’ Intermediary Guidelines and Digital Media Ethics Code have been created to address concerns like lack of transparency, disinformation and misuse across social media platforms in India, providing content regulation mechanisms.

In addition to these, considerations for Data Privacy and Protection are vital. Major standards like the EU General Data Protection Regulation (GDPR) and the APEC Cross-Border Privacy Rules System (CBPR) must be respected and considered in the Trust and Safety strategies of digital entities operating worldwide.

Law implication on social media: use case

Under the EU Digital Services Act (DSA), social media companies are obliged to submit their transparency reports, including the number of human moderators, error rates, and how quickly they comply with EU member states' requests, among many other factors. These documents aim to provide unprecedented insight into resources dedicated by platforms like Facebook, Snapchat, TikTok, and others for moderating illegal, hateful, or fraudulent content and general service statistics like user numbers. In early November 2023, major social media companies submitted their first reports and were mandated to repeat that reporting process every six months.

(Source: Le Monde)

5. Talent management

Finding the right talent for Trust and Safety is becoming challenging due to the need for a diverse skill set, relevant experience, and specific knowledge, especially when the demand for T&S services grows, contributing to the substantial talent shortage. As digital platforms expand and online interactions increase, the need for adept professionals in this field becomes even more pronounced, making recruiting and retaining skilled individuals an ongoing issue for various organisations.

Below are the most wanted positions in the industry:

- In this business, an exceptional T&S leader is crucial for navigating regulatory landscapes, fostering improvement, inspiring teams, supporting goal achievement, and strategically addressing emerging threats.
- In the tech context, an ideal T&S team must offer diverse capabilities, including infrastructure management, security programming expertise, data analytics, policy creation, and quality assurance.
- For content moderation, vital factors include language proficiency, emotional intelligence, cultural awareness, resilience under pressure, and handling exposure to harmful materials.

Trust and Safety services review

In light of the current state of Trust and Safety, its impending significance, and the increasing and more detailed obligations for more comprehensive user protection, higher standards, and greater transparency, one aspect is unquestionable – T&S will advance and become the top priority for a growing number of entities seeking competitive advantage in the evolving digital landscape.

In such demanding times, T&S will transform from being typically a game-changer for user satisfaction and retention to a must-have option for most social media platforms and other online services committed to operating online ethically and lawfully.

Consequently, companies – instead of answering the question "if"- must determine "how" to build a Trust and Safety strategy that secures users and preserves brand integrity optimally, efficiently and adequately. As the future will hold new responsibilities and capabilities, it is necessary to proactively prepare digital businesses to seamlessly allocate resources, integrate technologies and adapt processes for effective Trust and Safety management.

Further are suggested key components to consider when creating a modern T&S strategy. This list can serve as a "menu card" that offers a rich choice, broad possibilities and tailored solutions for specific challenges and objectives, which can be selected and mixed individually. With this approach, organisations have the flexibility to make informed decisions on customising their Trust and Safety framework. They can choose elements based on the unique demands of their digital environment, fostering adaptability, proactive risk mitigation, and sustained brand credibility.

1. User-generated content (UGC) moderation

Content moderation is a cornerstone of Trust and Safety, helping safeguard individuals and communities from offensive or illegal content that users commonly and increasingly share online in various forms, including posts, discussions, forum publications, comments, articles, multimedia or videos. To better illustrate the specificity and range of offensive or disruptive UGC materials, examples include:

- Using discriminatory language
- Endorsing violence or terrorism
- Fostering illegal actions like hacking
- Sharing threatening information
- Propagating misleading political content
- Publishing sexually explicit materials
- Provifing false financial or medical information
- Spreading sensationalist or clickbait content

Exposure to these and many other potential threats can be mitigated by monitoring, screening, and addressing the issues while relying on diverse content moderation methods, technologies, and contributors. Among the most popular and significant techniques are:

Pre-Moderation

Proactive manual review before content is published, preventing harmful materials from appearing.

Post Moderation

Reactive manual review after publication, supported by real-time tools, user reporting, and community guidelines.

Automated Moderation

Advanced tools and filters detecting and handling specific content.

Hybrid Moderation

Real-time automated screening with human involvement for complex issues, offering flexibility for diverse content analysis

Reactive Moderation	Users flagging or content reporting through tools, like report buttons or customer support tickets.
Distributed Moderation	User-driven moderation using rating and voting systems to elevate highly rated content while concealing or removing low-rated material.
Community-based Moderation	Active community participation in reporting, flagging, or rating content.

What is important to emphasise here is that one type of moderation is typically insufficient to handle threats entirely and successfully. Sometimes, it is necessary to combine diverse methods, especially when dealing with a comprehensive and high volume of content generated in a dynamic and vital online space where maintaining quality and safety is challenging through a single moderation approach.

Besides the need to address the most obvious vulnerabilities, the crucial factor is guaranteeing a thorough and customised strategy for content moderation. This involves clearly outlining what aligns with the brand's values and is deemed acceptable and what may be unacceptable, even if it is not immediately apparent. The same content can be managed differently on diverse platforms based on context and guideline specificity. For instance, a family-friendly web service emphasising children's safety may flag a picture of a child playing in a pool. At the same time, a more general social media site might not moderate the same content.

When done well, moderating services are instrumental in creating a safe space where people of different ages, genders, sexes, religions, ethnicities, nationalities, interests, or professions feel welcome, free of harm and more committed.

TikTok Use Case

On its website, TikTok showcases the approach to content moderation and shares its dedication to maintaining a safe and inclusive environment for its growing global community of creators. The company informs that, with over 40,000 safety professionals, TikTok relies on a combination of guidelines, automated moderation technology, and human moderators.

(Source: TikTok.com)

2. Content moderation complementary services

Nowadays, enhancing content moderation strategy with various valuable services and solutions that collectively contribute to a holistic approach is vital, addressing multiple aspects of content quality, safety, and compliance. These primarily include:

Establishing Community Standards:
This initiative defines acceptable behaviour and incorporates it into safety policies and guidelines for moderators and users. Making the rules easily accessible to the audience is critical for better understanding what is allowed and what is forbidden, promoting transparency and security of a digital platform.

X Use Case

X's rules have been designed to uphold user security and well-being. They are in place to ensure that everyone can safely participate in public conversation, contributing to the value of a vibrant and inclusive global discourse. They also inform about the prohibition of violence, harassment, and similar behaviours and aim to prevent actions that discourage individuals from freely expressing themselves.

(Source: X.com)

Quality Assurance: This involves ongoing evaluation, monitoring, and improvement of the content moderation processes to uphold consistency, accuracy, and effectiveness in handling user-generated content. The undertaking allows for driving agents' decision-making, streamlining moderation workflows and processes, ensuring consistency, providing moderators with feedback, and growing their skills accordingly.

UBISOFT Use Case

UBISOFT, a multinational video game company, enhanced its T&S strategy with a global Code of Conduct focusing on safety, respect, and inclusivity. It tackles toxic behaviour, deters cheating, and ensures accountability through investigations. Emphasising security, UBISOFT advises against sharing personal information to foster a safe and inclusive gaming environment, promoting fair play and player well-being.

(Source: Ubisoft.com)

Industry-Focused Moderation: This is exemplified by in-game moderation & monitoring, tailored to address the specific needs and dynamics characteristic of the gaming sector. It involves a deep understanding of player communities, their expectations, styles, behaviours, and needs. In such cases, moderators must be fluent in-game mechanics and adept at handling real-time interactions, ensuring a seamless and positive gaming experience.

Fraud Review & Investigation: This encompasses an in-depth review and assessment of potentially vulnerable content to detect and address fraudulent activities.

Fake Content Identification: This refers to utilising advanced image and video analysis techniques to detect fake content that is difficult to discover manually.

Ad Reviews & Compliance: This entails a detailed review of the advertisements to check if they comply with legal and ethical standards and address the violations when necessary.

Developer Compliance: This activity helps determine whether the developers adhere to platform policies and guidelines in their content creation while mitigating potential risks.

Content Tagging and Labelling: This is a way to enhance content classification, searchability, and safety through systematic content classification.

Identity & Account Authenticity: This involves verification and validation of user identities and the authenticity of accounts.

Content curation & updating: This allows for efficient management of user-generated content and adjustment of moderation practices to meet evolving user needs, technological changes, and digital landscape shifts, fostering innovation.

3. Trust and Safety compliance services

Trust and Safety compliance should ensure adherence to virtual space terms and conditions through various activities, such as UGC monitoring, risk mitigation strategies or efficient benchmarking. These can be supported with services, like:

Trust and Safety Compliance: These involve UGC monitoring to ensure platform terms and conditions adherence. The services encompass policy enforcement, including content removal and the adaptation of guidelines to address evolving challenges, thereby maintaining a secure and respectful online environment.

T&S Policy & Risk Advisory: It is about guiding and developing terms and conditions policies and risk management, advising on potential risks and offering strategic solutions to mitigate them.

Analytics & Benchmarking: This Trust and Safety initiative enables the assessment of the enforcement effectiveness of the Terms and Conditions through moderation outcomes analysis, benchmarking against industry standards, and leveraging data-driven insights for the policies enhancement.

4. Data, infrastructure and virtual assets protection

In the domain of safeguarding data, infrastructure, and virtual assets, a multifaceted T&S approach is essential for maintaining security and privacy, including:

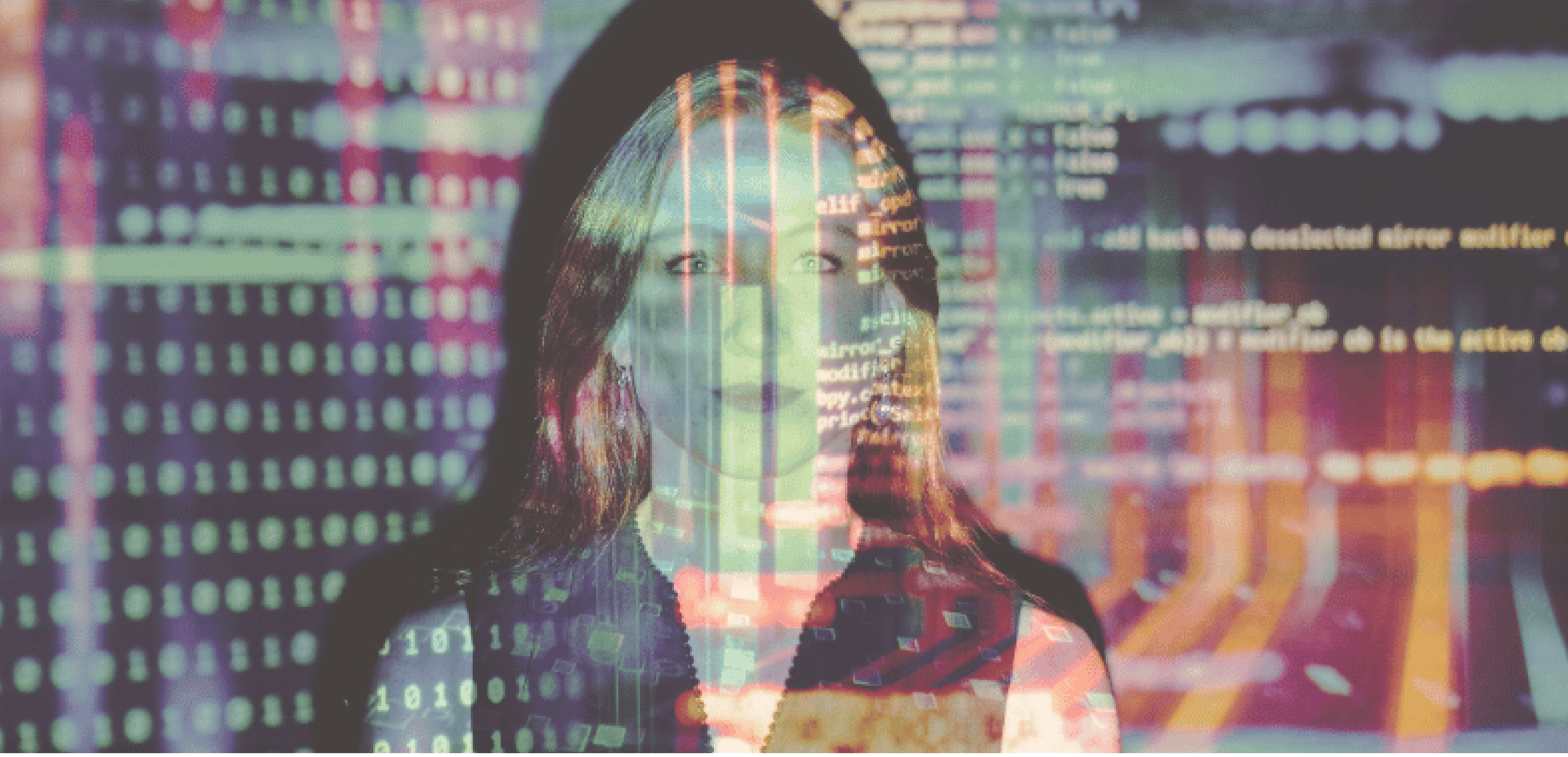
Privacy Protection Measures: This involves utilising safety measures to prevent unauthorised access, misuse, or theft of personal information collected, processed and stored by online businesses. Caution should also be exercised against manipulative or invasive data-driven profiling, sharing data with third parties, and excessive data collection. Maintaining privacy and freedom of speech and expression is another crucial factor when safeguarding personal information.

Cybersecurity Measures: This encompasses a set of solutions, tools and technologies to collectively create a secure digital environment by protecting users' sensitive information and preventing cyber-attacks. Among various examples are encryption protocols to secure the transmission of sensitive information and multi-factor authentication tools, adding an extra layer of security. It is also necessary to conduct regular security audits of the platform infrastructure and keep all systems, software, and applications current.

Protection of Virtual Assets: It refers to detecting and preventing unauthorised transactions, regularly reviewing user activity for suspicious behaviour, and securing the storage and transfer of virtual assets. These include digital wallets, blockchain technology, two-factor authentication, encryptions and monitoring.

5. Wellbeing & resilience programme development

Ensuring the wellbeing of content moderators is crucial for cultivating a positive work environment and mitigating the challenges associated with moderating sensitive content. This entails the development of comprehensive support systems, including resources, initiatives, and mechanisms aimed at providing assistance. Key components of these programs should include clear guidelines, counselling services, ongoing training, rotational shifts, peer support networks, regular check-ins, wellness workshops, anonymous reporting channels, health insurance benefits, and feedback loops. Together, these elements foster resilience and promote content moderation teams' mental and emotional wellbeing.



The future of Trust and Safety after the AI revolution

Predicting future trends in Trust and Safety content moderation involves anticipating how technology, user behaviour, and regulatory landscapes may evolve. This is what one can expect in 2024:

1. Artificial intelligence in action

As we look ahead, AI-driven automation is set to continually revolutionise Trust and Safety, making it even more adaptable, comprehensive and efficient when navigating the changing landscape of increasingly sophisticated online threats. Such solutions as machine learning, deep learning, and natural language processing will continue to enhance the accuracy and capability of threat identification, enabling real-time issue addressing and compliance monitoring and enforcement. Artificial intelligence will also be a key driver of personalised moderation, ensuring that content remains within the bounds of community guidelines and standards to a larger extent. Going further, leveraging predictive analytics to anticipate potential risks will be increasingly common and more efficient, preventing safety issues before they escalate. T&S teams will also prepare for technological advancements, such as quantum computing, by understanding, researching and implementing quantum-safe security measures to protect user data.

2. Balancing AI and human intelligence

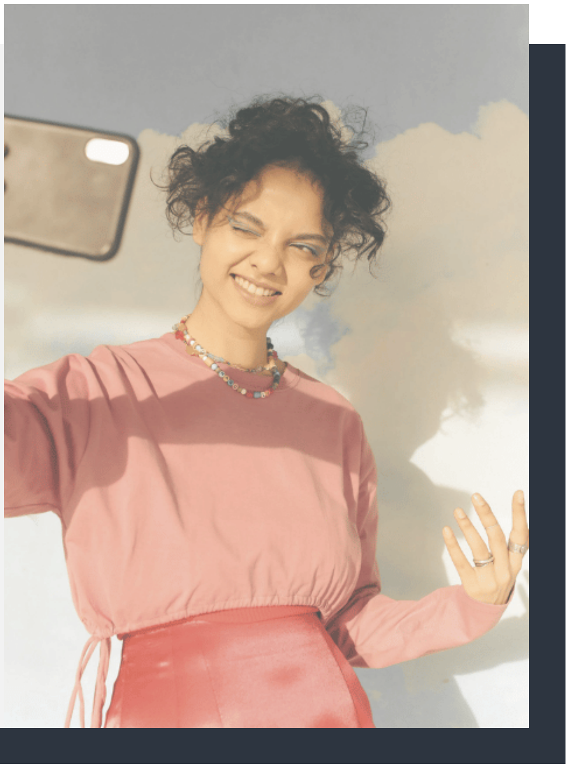
Soon, a notable trend in Trust and Safety will also involve the synergistic collaboration between AI-empowered technology and human agents. Although artificial intelligence brings immense opportunity for T&S enhancement beyond human capacities, such as automating processes, enhancing scalability, and greater accuracy, the human touch will remain invaluable and irreplaceable, providing empathy, cultural sensitivity, and a deep understanding of cultural nuances. This is still a human advantage over AI regarding the ability to interpret highly subtle sarcasm and humour and handle content that does not neatly conform to predefined rules. Combining human skills with innovative solutions, one can create the winning T&S strategy, amplifying the effectiveness of the overall T&S process and forming a well-balanced partnership carefully tailored to specific circumstances.

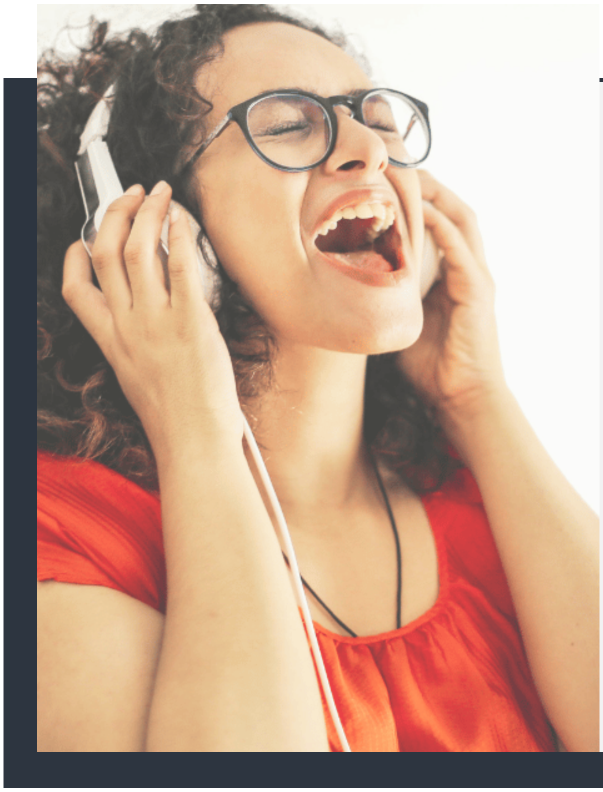
Recognising the importance of cultural context in global content moderation, skilled and knowledgeable human agents play a crucial role in nuanced decision-making, understanding community-specific norms, and grasping subtle language or contextual nuances, which automated content moderation systems might overlook or misinterpret, potentially leading to inappropriate content being allowed or deleted unnecessarily. Such considerations are essential to prevent misunderstandings and offence, especially across languages and diverse cultures.

Facebook Use Case

Facebook is an example of a social media platform using artificial intelligence in content moderation. AI plays a central role in detecting and removing content that violates standards, often before user reports. Content is sent to the Facebook human review teams distributed worldwide when necessary. These reviewers focus on the most harmful content, ensuring a balance between AI efficiency and human decision-making to maintain a safe and expressive environment.

(Source: Facebook.com)





YouTube Use Case

During the pandemic, YouTube relied more on machine moderators to filter content, but this approach led to excessive removal of videos, including many that did not violate any rules. As a result, YouTube has reverted to using more human moderators to address the issue. These shed light on the relationship between human moderators and artificial intelligence systems

(Source: Financial Times)

3. Advancing content moderation challenges

What else will grow in significance is a shift toward more comprehensive content, particularly video and multimedia materials, which will also be significant in the Future Trust and Safety, demanding more sophisticated and time-intensive moderation processes. Overcoming this challenge involves leveraging emerging technologies, expanding moderation teams, or encouraging increased user reporting. Another key issue for digital businesses will be a heightened focus on child safety, incorporating advanced automation alongside user-friendly reporting tools, age restrictions, educational initiatives, and collaborative efforts with child protection organisations.

Key Resources

1. Cybercrime Magazine, 2023 Cybersecurity Almanac: 100 Facts, Figures, Predictions, And Statistics.
2. Anti-Defamation League (ADL), Online Hate and Harassment Reaches Record Highs, ADL Survey Finds.
3. Microsoft, Microsoft Entra: 5 identity priorities for 2023.
4. McKinsey, New survey reveals $2 trillion market opportunity for cybersecurity technology and service providers.
5. The Business Research Company, Content Moderation Solutions Global Market Report 2023.
6. European Union Agency for Cybersecurity (ENISA), NIS INVESTMENTS Cybersecurity Policy Assessment (November 2023).
7. PwC, 2022 Global Digital Trust Insights.
8. McKinsey, The consumer-data opportunity and the privacy imperative taking a thoughtful approach to data management (2020).
9. UK Government Service, A guide to the Online Safety Bill.
10. Singapore Statutes Online, Online Safety (MISCELLANEOUS AMENDMENTS) ACT 2022.
11. Family Online Safety Institute, What You Should Know About the Kids Online Safety Act
12. California's Age-Appropriate Design Code Act
13. European Commission, Shaping Europe's digital future
14. Le Monde, Content moderation: Key facts to learn from Facebook, Instagram, X and TikTok transparency reports
15. Financial Times, YouTube reverts to human moderators in the fight against misinformation.
16. Brand website: TikTok, Facebook, YouTube, UBISOFT.

A few words about Conectys

Conectys is a BPO vendor with an industry focus, a partnership mindset, and the right size for the international expansion of clients' brands.

Trust and Safety

We offer human-first global Trust and Safety solutions to deliver 24/7/365 protection from evolving regulations and user harm.

Customer Experience

We provide customer experience management and global contact center solutions for international brands.

Global Outsourcing Partner

Our assets are: 14 locations, 35+ languages, WFH agents, and 24/7 availability.

Let's talk!

Contact us

US - 1469 532 0215 UK - 44 203 318 1593 EU - 32 929 8011 HK - 852 800 930 130

sales@conectys.com www.conectys.com