



TRUST & SAFETY IN TRANSITION

- ● TRENDS, CHALLENGES,
AND FUTURE INNOVATIONS



Table of Content

- 1** Introduction
- 2** A State of Digital Security: Strengths and Weaknesses
- 3** Key Challenges in Building the Impactful Trust and Safety Framework
- 4** Efficient Trust and Safety Strategies That Make an Impact
- 5** A Role of Agentic AI in Trust and Safety
- 6** AI and Human Working Harmoniously
- 7** Conclusion
- 8** A Few Words About Conectys
- 9** Contact Us

Introduction

The digital world connects billions of people, enabling various activities, from social interactions and entertainment to education, financial management, work, and many more. However, as this online population grows and interactions multiply, cracks in its foundation are becoming increasingly apparent. Harassment, hate speech, and cybercrime have become as pervasive as the platforms themselves.

This presents a significant challenge: moving beyond what has traditionally been considered efficient in Trust and Safety, an initiative that serves as the backbone of the virtual ecosystem, ensuring its integrity and fostering secure, inclusive user experiences.

Today, it is about creating intelligent, resilient, and forward-thinking systems that stay ahead of evolving threats and effectively safeguard vulnerable users, virtual assets, data privacy, community integrity, and brand reputation.

Achieving this requires integrating innovative technologies, with artificial intelligence leading the charge. The focus must shift from reactive measures to a proactive, data-driven approach embedded seamlessly within platform processes—especially those related to content moderation, compliance, and cyber threat mitigation.

Strengthening the role of **Data Annotation and Labelling** is also paramount. As a foundational element, it maximises AI's potential by supplying algorithms with high-quality, accurate datasets, enhancing their ability to detect and prevent violations.

Ultimately, the synergy between human expertise and advanced technology is a cornerstone. Complementing each other's unique strengths, they contribute to a dynamic and adaptive protective strategy—ensuring online spaces remain safe, inclusive, and conducive to positive interactions.

A State of Digital Security: Strengths and Weaknesses

Today's cyberspace is a dynamic, rapidly evolving environment that is both crowded and multicultural. As it grows, it becomes an ever more complex landscape where individuals worldwide interact, communicate, and share content in unprecedented ways. But, while this exciting sphere offers remarkable opportunities for connection and innovation, it also brings new and increasingly serious challenges and risks.

Exploring the scale of the global digital presence, as of 2024, around 5.5 billion people worldwide used the internet, up from 5.3 billion the year before, meaning 68 per cent of the global population was online, according to Statista. Additionally, other Statista data highlights the range and impact of the cyber footprint people leave each day, whether anonymously or personally, by sharing thoughts, opinions, pictures, or videos, for example:



In 2024, **5.22 billion people**, 63.8% of the global population, used social media—a figure led by China, India, the US, and Europe, expected to surpass six billion by 2028.



In 2024, internet users spent an average of **143 minutes daily** on social media, commenting on posts and sharing or liking updates, photos, and videos as their preferred activities.



The digital footprint has grown significantly, exemplified by Tripadvisor, which surpassed **one billion reviews** in 2021 and covered over **eight million travel listings** by 2023.

The numbers underscore how the virtual realm has evolved into an integral part of our lives, offering unparalleled experiences and bringing together individuals from diverse cultures, locations, and social backgrounds. Yet, as it grows increasingly crowded and heterogeneous, it also mirrors the dynamics of the offline reality—where fair play and sincerity coexist with dishonesty—creating both challenges and opportunities for constructive and harmful behaviour. The latter is what the Anti-Defamation League (ADL) regularly assesses, alongside many other organisations evaluating the severity of online harm. According to some of the ADL's findings:

52%

In 2023, 52% of Americans experienced online harassment, a 12% increase from the previous year.

51%

Teenage cases surged from 36% in 2022 to 51% in 2023, with severe incidents nearly doubling.

75%

Around 75% of the US gamers reported exposure to harassment or hate, highlighting the pervasive nature of online harm.

In addition to the growing trend of virtual abuse, there has also been a significant expansion in cybercrime activities involving much more sophisticated threats such as data breaches, ransomware, malware, phishing, social engineering, identity fraud, digital espionage, and crypto crime. These malicious practices cause critical damage, including theft of data or intellectual property, hidden costs such as money spent on restoration efforts, direct financial losses, and significant reputational harm.

According to Cybersecurity Ventures, global cybercrime costs are projected to reach \$10.5 trillion annually by 2025, up from \$3 trillion in 2015. This represents the largest transfer of economic wealth in history.

These alarming trends highlight the urgent need for stronger online security efforts. However, ensuring adequate protection requires more than traditional Trust and Safety approaches. To address complex demands, emerging threats, and stricter obligations, organisations must move beyond what was once considered sufficient, adopting relevant strategies and solutions in the face of evolving challenges.

Key Challenges in Building the Impactful Trust and Safety Framework

Emerging trends in Trust and Safety are essential. They guide expectations, enhance capabilities, and help businesses stay ahead of ever-evolving challenges. With online spaces becoming more dynamic and user-driven, doing nothing or doing less is no longer an option. Companies and communities must keep pace with the novelties to uphold security, fairness, and credibility.

1. Technological Progress

Technological advancements challenge digital entities to continuously adapt, balancing efficiency, productivity, and protection while addressing the evolving nature of online risks and meeting regulatory demands.

In the initial stage, monitoring and analysing emerging innovations is essential to evaluate how effectively they align with strategic objectives. This involves weighing the pros and cons, determining when an investment will yield tangible benefits, and identifying scenarios where adopting a solution may not be the optimal choice. Understanding these nuances ensures that resources are allocated wisely, focusing on technologies that enhance Trust and Safety efforts while avoiding unnecessary complexity.



Key technologies shaping T&S include artificial intelligence, natural language processing, machine learning, data labelling and annotation tools, various cybersecurity tools, automated moderation systems, image and video recognition algorithms, blockchain, real-time risk monitoring programs, compliance software, and advanced user authentication solutions.

Among all, AI-driven algorithms lead the way in revolutionising Trust and Safety, significantly enhancing the efficiency and accuracy of digital protective measures. They enable real-time detection of harmful content, combat deepfakes with precision, and identify threats such as fraudulent activity or phishing attempts at scale. However, their deployment is not without obstacles. If left unaddressed, risks like algorithmic bias and ethical concerns, including data privacy, could undermine user trust. Moreover, AI tools require rigorous data labelling, ongoing training, skilled oversight, and calibration to ensure fairness and effectiveness. As threats become increasingly sophisticated, blending AI capabilities with human expertise remains essential for context-sensitive and ethical decision-making in this ever-evolving landscape.

For instance, **Generative AI** enables machines to create content like text, images, music, and videos by learning patterns from large datasets. While this offers exciting possibilities, it also raises Trust and Safety concerns. The key lies in AI training designed to mitigate these risks:



One major concern is **AI hallucination**, where the model generates inaccurate content due to its inability to grasp nuances, cultural context, or language intricacies.



Additionally, **deepfakes** created by generative AI—highly realistic and deceptive—can spread misinformation and harm reputations if left undetected.



Another issue is the potential for **oversharing sensitive data** during training, which cybercriminals could exploit for malicious purposes.

Ultimately, when embarking 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. The Explosion of User-generated Content and Its Growing Dark Side

The rise of user-generated content (UGC) has revolutionised digital engagement, with millions of posts, photos, and videos shared on social platforms continually. The statistics provided by LocalIQ highlight this phenomenon:

Every day, Instagram users share **95 million** photos and videos, which equals **66,000 every minute.**

Facebook users post **510,000** comments, **293,000** status updates, and **240,000** photos every **60 seconds.**

Snapchat sees **5 billion** Snaps created daily worldwide, while **350,000** tweets are sent every minute on X (formerly Twitter).

While UGC is a powerful asset for virtual spaces, creating unparalleled opportunities for promotion, popularity and retention, it also presents hurdles. Without proper oversight, such materials can become a vehicle for harm or misinformation—whether intentional or accidental—undermining trust and compromising platform reliability.

Mitigating the risks requires a strategic approach that combines advanced machine learning algorithms for efficiency and scalability with skilled human moderators for nuanced oversight. 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's specific regulations, user groups, preferences, demographics, languages spoken, or cultural backgrounds.

3. The Advancing Landscape of Cyber Threats

In the early days of the internet, when online presence was limited and web activity less complex, security incidents were infrequent, their impact was smaller, and attacks were typically human-driven and straightforward. The scope of potential harm was narrow, and security measures were simpler to implement.

Cyber threats have become a persistent, global challenge as digital technology permeates. They have grown intricate, often orchestrated by complex networks and bolstered by AI capabilities. They result in more severe and far-reaching consequences, exploiting vulnerabilities with alarming precision and scale.

Traditional security methods can no longer keep pace, forcing organisations to adopt advanced detection systems, upgrade cybersecurity infrastructure, and provide ongoing employee training—all of which demand significant investment. The emergence of quantum computing, which poses a potential threat to existing encryption methods, compounds these challenges. As the capabilities advance, the urgency to develop and implement quantum-resistant security measures surges, requiring businesses to address these risks proactively.



4. Navigating the Complexities of Regulatory Compliance

Regulatory compliance presents another formidable challenge to Trust and Safety. Global legislative initiatives place importance on enhanced privacy, user protection, and transparency while holding online platforms accountable for maintaining safe, abuse-free environments. For instance, regions such as the EU, UK, US, Singapore, and India have recently introduced stringent regulations to adapt to emerging technologies.

These laws require substantial investments in legal expertise, technology upgrades, and operational adjustments. Failure to comply can result in severe penalties and reputational damage, making it crucial for companies to navigate diverse regulatory landscapes proactively.

5. Impact of the Gig Economy and Remote Work

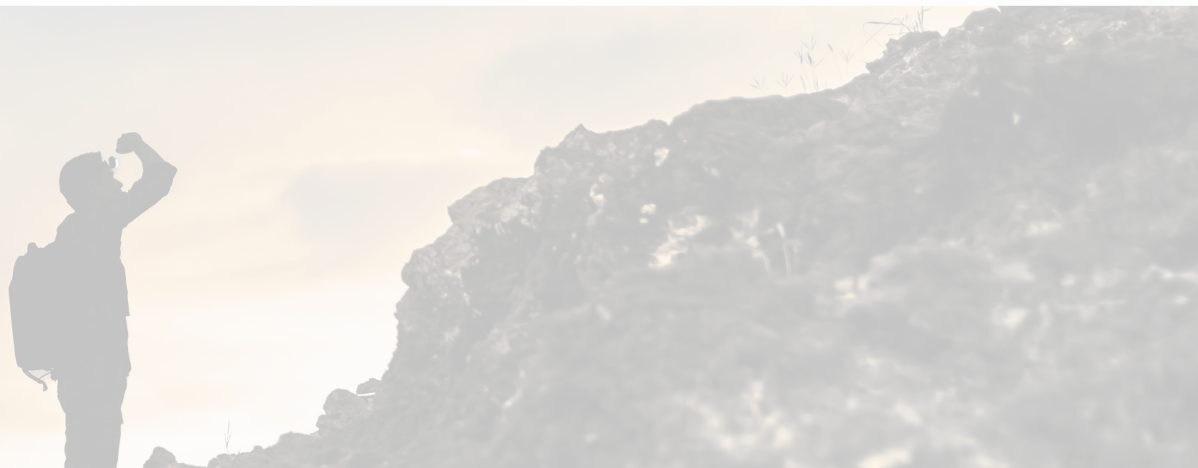
The rise of the gig economy and the shift to remote work have additionally influenced Trust and Safety dynamics. With tasks increasingly performed on digital platforms or outside traditional office settings, new dangers arise, including data breaches, industrial espionage, and the mishandling of sensitive information. These trends necessitate robust security measures, clear guidelines, and transparent processes to ensure accountability across a geographically dispersed, often borderless workforce. Addressing technical limitations, external pressures, and internal constraints requires a proactive, problem-solving approach to maintain a safe and secure environment.

6. Tackling Talent Shortages and Employee Well-being

The growing demand for skilled Trust and Safety professionals—particularly in cybersecurity, data analytics, regulatory compliance, and **content moderation**—has created significant talent shortages. The competitive hiring landscape and the need for specialised skills make recruiting and retaining workers a pressing issue. Additionally, content moderators often encounter distressing material, which can negatively impact mental health. Organisations must prioritise wellness programs, provide mental health support systems, and adopt advanced recruitment and training methods to ensure employee well-being and sustained performance.

7. Aligning with Privacy, Sustainability, and ESG Principles

As user expectations evolve toward higher standards, virtual platforms need to concentrate accordingly on enhanced data privacy, sustainability, and ESG principles. Customers increasingly demand that digital firms demonstrate ethical governance, environmental responsibility, and inclusive practices alongside delivering functional services. Balancing these demands with profitability requires a careful mix of innovation, transparency, and accountability.



Efficient Trust and Safety Strategies That Make an Impact

The Trust and Safety initiative is a critical priority for online businesses, especially as user protection, higher standards, and transparency become paramount for ethical and lawful operations. This involves preparing to allocate resources, integrate technologies, and adapt processes for seamless T&S management.

The following components provide a flexible approach to building a modern T&S strategy, offering tailored solutions that align with specific challenges, risk mitigation needs, and brand goals. Companies can enhance their adaptability, reputation, and overall safety by selecting the right mix of elements:

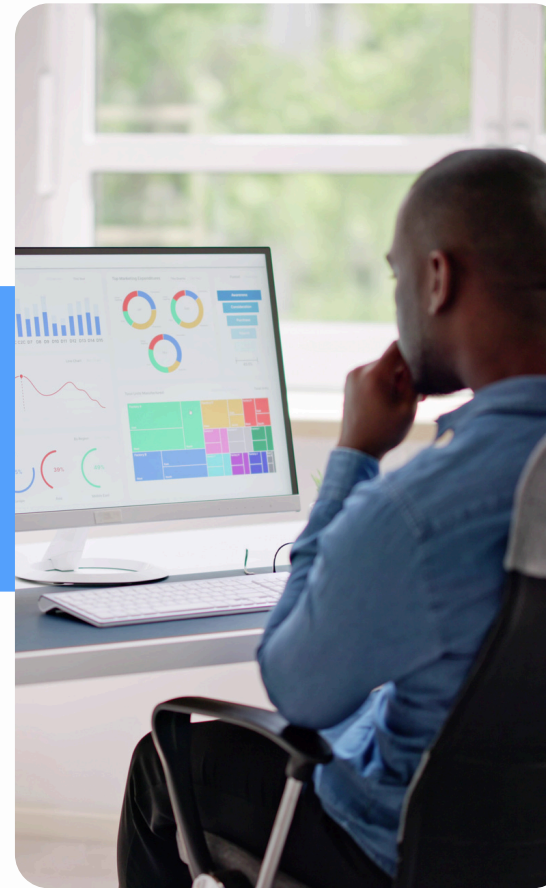
1. Data Labelling and Annotation

Data labelling and annotation are key technologies that enhance the quality, diversity, and precision of data used in AI and machine learning projects. These processes involve tagging or categorising raw records to make them usable for training AI models, enabling them to recognise patterns and make accurate predictions. Ultimately, data labelling and annotation solutions are revolutionising Trust and Safety, helping online services to safeguard digital environments better.

With high-quality, annotated data, AI models are well-trained to detect harmful material such as hate speech, misinformation, and abusive behaviour accurately. This enhances the safety of virtual spaces, supports compliance with regulations, ensures adherence to standards, and protects visitors from exposure to damaging content. Ultimately, data labelling and annotation are critical enablers for fairness, inclusivity, and transparency in AI-powered systems, given their role in reducing biases and ensuring ethical AI deployment.

Key data labelling and annotation features include:

- **AI-Assisted Annotation:** AI-powered tools automate the annotation process by pre-annotating data, reducing manual effort by up to 90%. This allows T&S teams to focus on auditing and refining annotations, ensuring faster and more scalable results. Examples include machine learning (ML) pre-annotation and auto-correction tools that enhance efficiency without compromising quality.
- **Multi-Format Data Support:** Seamless annotation across diverse data types, including text images, audio, video, 3D models, and sensor data, ensures versatility for a wide range of AI applications. Supported formats such as JSON, PCD, CSV, and TXT enable easy integration with downstream AI workflows, providing compatibility for specialised use cases.



- **Video Annotation:** Video content is increasingly critical in Trust and Safety and other applications. Features like frame-level scene classification, object tracking, and occlusion handling ensure accurate annotations, even for complex scenarios. Real-time annotation capabilities are essential for applications requiring live monitoring or smarter threat detection.
- **Data QA and Real-Time Feedback:** Integrated quality assurance tools enable real-time communication between annotators and managers. This ensures immediate error correction and maintains high data accuracy for AI training. Iterative feedback loops enhance long-term model performance, creating a continuous improvement cycle.
- **Workforce Management:** Advanced tools for managing internal and external teams streamline task distribution, monitor performance, and ensure consistent output quality. Performance tracking metrics such as time-to-annotate, accuracy rates, and task completion timelines provide actionable insights for optimising team efficiency.
- **Integrated Labelling Services:** Platforms offering access to professional annotators for specialised tasks enhance scalability and expertise without overburdening internal resources. This includes support for annotation at scale, focusing on niche areas like 3D cuboid annotation, LiDAR optimisation, and polynomial segmentation for industries like autonomous vehicles, robotics, and smart cities.
- **Ethical Annotation Practices:** Ethical considerations are integral to ensuring datasets are inclusive, unbiased, and aligned with global standards like GDPR and the EU AI Act. Annotators follow strict protocols to promote fairness, inclusivity, and transparency, helping build AI systems that reflect diverse perspectives and avoid reinforcing biases.
- **Global Reach and Local Expertise:** With a network spanning multiple continents, platforms can provide culturally nuanced and language-specific annotations. This global reach, combined with local expertise, ensures high-quality results tailored to regional needs, which is critical for NLP and localisation projects.
- **Security and Compliance:** Adherence to enterprise-grade standards like GDPR, ISO 27001, and SOC 2 ensures the highest data security and privacy level. Features like encryption, two-factor authentication (2FA), and role-based access controls (RBAC) safeguard sensitive datasets throughout annotation.

2. Data Focus Strategies

Data-driven strategies are at the core of effective Trust and Safety initiatives. They enable businesses to make informed decisions and build a strong, responsive Trust and Safety framework for addressing evolving digital landscape challenges. These encompass:

- **Real-time Data Collection:** Gathering data from various sources, such as user interactions, client feedback, and content moderation activities, provides a foundation for generating actionable insights.

- **Advanced Analytics and BI Solutions:** Leveraging cutting-edge Data Analytics and Business Intelligence tools enables organisations to enhance operational efficiency, mitigate risks, and maintain compliance with evolving standards.
- **Comprehensive Data Management:** Ensuring data integrity through meticulous labelling, curation, and automation optimises processes and enhances reliability.
- **Streamlined ETL and Real-time Analytics:** Efficient Extract, Transform, Load (ETL) workflows, combined with real-time analytics, deliver timely insights that drive informed decision-making, accelerate AI development, and refine model performance.

3. Content Moderation: Adapting to Complexity

Content moderation practices have transformed to address the scale, variety, and speed of online interactions in increasingly complex virtual spaces. Traditional approaches are no longer sufficient, and modern strategies now prioritise proactive, adaptive solutions to manage text, images, videos, and live streams across diverse, multilingual platforms.

Key elements of impactful content moderation include:

- **Scalable Oversight and Harm Prevention:** By integrating artificial intelligence algorithms, moderation efforts become scalable and capable of predictive harm prevention. Advanced AI technologies, such as automated triaging, real-time learning models, and sentiment analysis, greatly enhance efficiency and accuracy.
- **Balanced Human-AI Collaboration:** Platforms can achieve context-aware moderation at scale by combining AI's precision with human judgment. Humans are fluent in recognising nuances and understanding cultural context—areas where AI may struggle. Meanwhile, AI offers automation and precision, empowering moderators with specialised training and robust knowledge management for higher productivity.
- **Customisable Frameworks:** Flexible moderation systems allow virtual entities to tailor policies, workflows, and AI models to meet specific needs and regulatory requirements. This adaptability ensures effectiveness across diverse digital landscapes and addresses unique audience and content challenges.
- **Oversight Practices and Adaptation:** Moderation systems must evolve to keep pace with shifting user behaviours, emerging threats, and changing regulatory landscapes. Continuous refinement ensures moderation practices remain resilient, relevant, and aligned with community expectations, creating a safer and more trustworthy online environment.



- **Empowering Users Through Education:** Educating visitors on responsible online habits is essential for fostering safer digital spaces. This proactive approach encourages individuals to participate in creating positive interactions, enhancing the overall effectiveness of moderation efforts.
- **Expertise and Advanced Tools:** Collaboration with external experts and adoption of AI-enhanced tools significantly strengthen oversight capabilities. These strategies are particularly effective for detecting and addressing specialised content challenges, such as hate speech, misinformation, or deepfakes, ensuring a comprehensive approach to maintaining digital trust and safety.

○ Refinement of Moderation Strategies

It is worth emphasising that many established moderation methods remain highly effective in addressing diverse threats, particularly when accuracy, flexibility, and scalability are paramount in tackling complex or sophisticated issues. While relying on a single approach can work in certain scenarios, combining multiple strategies and resources enhances overall effectiveness.

Key approaches include:

- **Proactive Reviews:** Manual reviews conducted before publishing content can prevent harmful material from reaching users.
- **Reactive Moderation:** This method addresses issues as they arise and is supported by real-time tools and user reporting systems.
- **Advanced Filtering:** Utilising AI-driven algorithms to detect and manage specific content types efficiently.
- **Integrated Tools:** Embedding immediate automated screening alongside user flagging tools or rating systems ensures a safe and trustworthy environment.

Moderation capabilities will continue to advance, propelled by sophisticated AI and strategic integration of AI with human expertise. This includes automated triaging to prioritise content for review, real-time learning models that adapt to new risks, and sentiment analysis that interprets tone and context, leading to more accurate moderation decisions.



Facebook Use Case

Facebook is an example of a social media platform using artificial intelligence in moderation. It 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 and 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)

4. 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:

- **Fraud Review and Community Standards:** Establishing clear community standards is essential for transparency and security. These guidelines should define acceptable behaviour and be easily accessible to moderators and users. Proactively addressing fraudulent activities through advanced analytics and manual reviews builds trust within the platform.
- **Quality Assurance:** Continuous evaluation and improvement of moderation processes ensure consistency, accuracy, and effectiveness. Regular feedback, specialised training, and streamlined workflows empower moderators to make informed decisions while enhancing their skills.
- **Industry-Specific Moderation:** Tailored approaches, such as in-game moderation, address the unique needs of specific sectors. For instance, moderators familiar with gaming dynamics can handle real-time interactions effectively, ensuring seamless experiences while tackling fraud or inappropriate activity.
- **Advanced Detection and Classification:** Modern tools like AI-powered image and video analysis, NLP, and sentiment analysis enhance the detection of fake content, hate speech, or deepfakes. Systematic tagging and labelling improve content organisation and searchability while bolstering safety measures.
- **Compliance and Accountability:** Rigorous ad reviews ensure alignment with legal and ethical standards, while developer compliance monitoring mitigates risks by enforcing platform policies. Identity verification and account authenticity checks further enhance platform security.
- **Adaptability and Engagement:** Moderation practices must evolve with user interactions, technological advancements, and regulatory changes. Regular policy updates and content curation keep platforms resilient and aligned with community expectations. User education, feedback, and collaboration with external experts foster safer digital spaces.

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)

5. Trust and Safety Compliance Services

Trust and Safety compliance ensures adherence to platform terms and conditions through activities like UGC monitoring, risk mitigation, and benchmarking. The most pivotal methods and solutions entail the following:

- **Policy Enforcement:** Monitoring UGC to ensure compliance with platform guidelines, removing harmful content, and adapting policies to address emerging challenges.
- **Policy and Risk Advisory:** Developing and refining terms and conditions while advising on potential risks and mitigation strategies.
- **Analytics and Benchmarking:** Assessing moderation effectiveness by analysing outcomes, benchmarking against industry standards, and using data-driven insights to enhance policies.
- **Proactive Threat Detection:** Identifying and addressing emerging risks, such as evolving harmful behaviour patterns, to pre-empt potential violations.
- **Transparency and Reporting:** Providing regular updates on compliance metrics, policy changes, and risk management strategies to foster trust and accountability.
- **User Support and Education:** Educating users about platform policies, empowering them to report violations, and fostering an informed community.
- **Regulatory Adaptability:** Staying aligned with evolving laws and global standards to ensure continued compliance and operational integrity.

Effective T&S compliance entails much more than simply enforcing platform rules—it is about cultivating a secure, inclusive, and dynamic digital space that evolves in response to emerging threats and user needs. With ever-growing volumes of content and sophisticated exploitation methods, such as fake accounts, deepfakes, and cyberattacks, platforms must continuously adapt their policies and procedures to stay ahead of these risks.

Identifying and preventing potentially harmful content before it reaches the community is crucial to sustaining user engagement and upholding the platform's reputation. It creates a dynamic balance between protecting users from harm and empowering them to express themselves within safe boundaries.

6. Protective Measures

Ensuring data security while preventing unauthorised access is a crucial objective for T&S, enabling organisations to stay ahead of user needs and expectations. Whether it involves personal data, sensitive information, or digital assets, each area requires robust measures to safeguard against external and internal threats.

Key measures to achieve this include:



Privacy Protection Measures prevent illegal access, misuse, or theft of personal information collected, processed, and stored by online businesses.

Careful attention should be paid to avoiding manipulative or invasive data profiling, sharing data with third parties, and excessive data collection. It's equally important to safeguard personal information while upholding the right to freedom of speech and expression.



Cybersecurity Measures encompass a variety of tools, solutions, and technologies designed to create a secure domain by protecting sensitive information and preventing cyber-attacks.

Encryption protocols secure the transmission of sensitive data, while multi-factor authentication adds an extra layer of security. Regular security audits of platform infrastructure and updating all systems are vital for protecting against vulnerabilities.



Protection of Virtual Assets involves detecting and preventing unauthorised transactions, regularly reviewing user activity for suspicious behavior, and securing the storage and transfer of digital assets.

Technologies like e-wallets, blockchain, two-factor authentication, and encryption are used alongside continuous monitoring to ensure secure and compliant handling of virtual assets.

7. Enhancement through Agile Scaling Gig Workforce Platforms

To optimise Trust and Safety operations, companies should harness the power of agile gig workforce platforms, which provide access to pre-vetted gig workers from 180+ countries and 80+ languages. These services allow for the rapid scaling of operations in response to fluctuating demand while ensuring that cultural nuances and regional specifics are recognised and addressed.

A key issue is leveraging AI-driven task allocation and real-time productivity monitoring to ensure that required measures are executed precisely and efficiently. This enables businesses to address high-volume periods, such as major events or crises, while maintaining quality standards and seamlessly integrating existing workflows.



8. T&S Teams Well-being & Resilience

The approach should prioritise the well-being of Trust and Safety agents, recognising the unique pressures they face. Their physical, emotional, and mental health support is crucial for their welfare and work effectiveness. A comprehensive program should span the entire employee journey, from recruitment to post-employment. This includes gamified onboarding resilience training, focusing on mental health, coping skills, and socialisation, with tailored support for varying needs. Workplace counselling should offer a confidential, non-judgmental space to address personal or work-related challenges, complemented by a 24/7 support system in multiple languages. Additionally, group interventions—such as psychoeducational workshops, creative activities, and team bonding exercises—can enhance mental health literacy, coping strategies, and internal relationships.

A Role of Agentic AI in Trust and Safety



Agentic AI, a sophisticated artificial intelligence capable of autonomously planning and executing tasks to achieve specific goals, is about revolutionising the Trust and Safety landscape. Its potential lies in drastically improving operational efficiency, adaptiveness, and proactive risk management. By functioning without direct human intervention, agentic AI is poised to take on the increasing demands of content moderation, data analysis, and emerging threats, providing a more efficient and scalable approach to safeguarding digital environments.

One of Agentic's primary advantages is its ability to automate content moderation at scale. In platforms with high volumes of user-generated content, such as social media networks or online marketplaces, agentic AI can swiftly process vast amounts of data, flagging harmful content like hate speech or spam in real-time. Its adaptive and contextual nature further strengthens moderation efforts by handling routine tasks autonomously, while more nuanced cases are escalated to human moderators for thoughtful review.

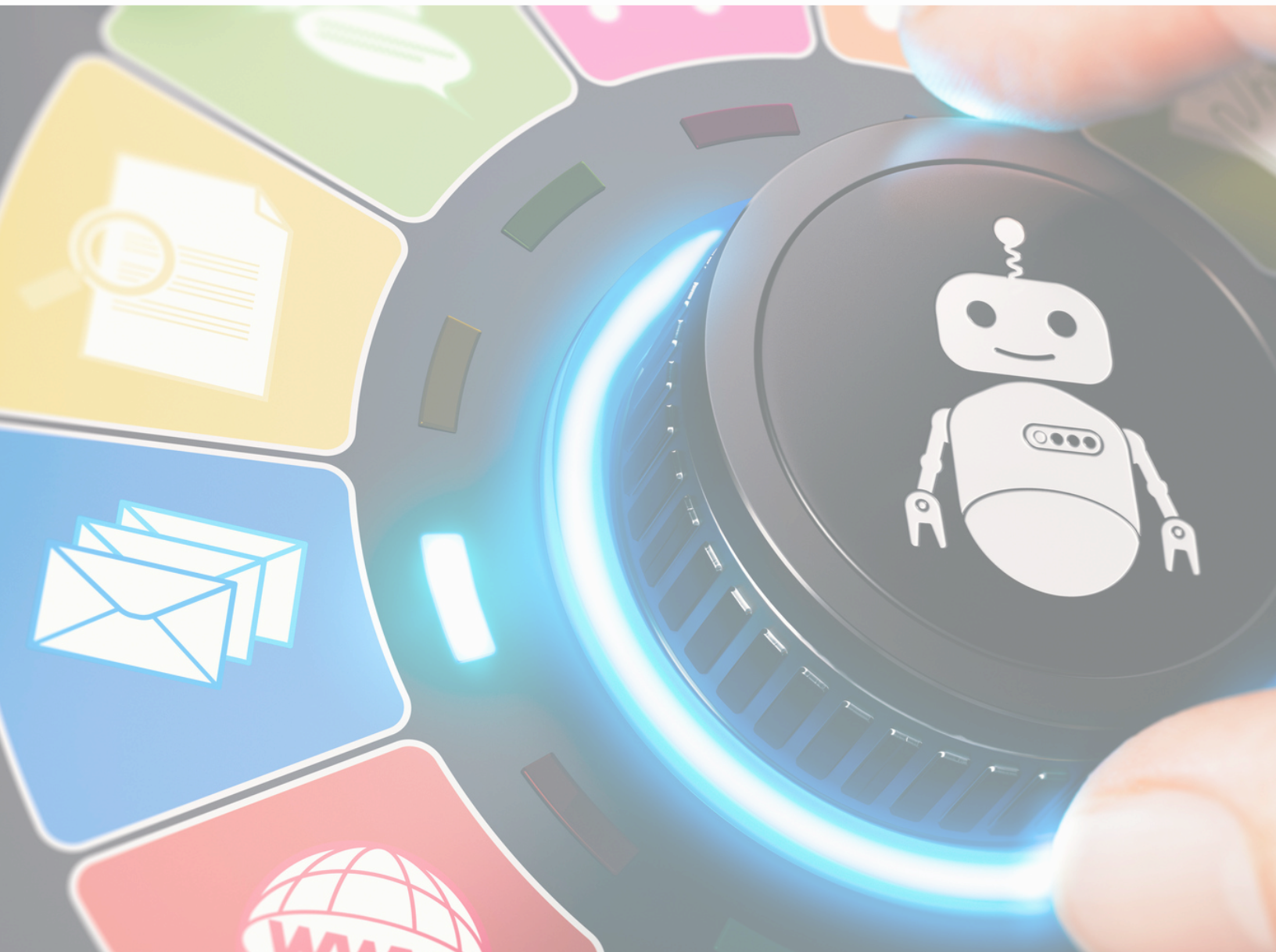
Additionally, Agentic's AI ensures compliance with evolving regulations by automating the assessment of content and user activities against legal and ethical standards. It streamlines audit trails, tracks compliance metrics, and supports transparent reporting—a critical capability for many entities operating in diverse jurisdictions. Technology also plays a pivotal role in handling crises. Whether it is a coordinated disinformation campaign or an unexpected surge in harmful content, Agentic AI can quickly scale operations, prioritise critical incidents, and allocate resources effectively.

Beyond content moderation, agentic AI plays a crucial role in proactive risk mitigation and personalised user experiences. By analysing patterns and detecting emerging risks early, AI can enable T&S teams to take preventative actions, reducing the likelihood of incidents and breaches. Furthermore, it can customise CX based on visitor profiles, tailoring content and interactions while maintaining rigorous safety standards.

AI and Human Working Harmoniously

While AI technology offers immense opportunities to enhance Trust and Safety processes—automating workflows, processing vast amounts of data, and identifying patterns or risks with remarkable accuracy and precision—it is not a standalone solution. The human touch remains invaluable, offering empathy, cultural sensitivity, and nuanced judgment that AI cannot replicate. For instance, humans excel at moderation, recognising sarcasm, humour, and social dynamics while effectively managing scenarios outside predefined rules or algorithms.

Companies can develop a robust and adaptive T&S strategy by blending human skills with AI-driven technology. At the same time, agents provide oversight to ensure fairness and emotional intelligence in content moderation, threat detection, and policy enforcement. Together, this collaboration enhances safety, responsiveness, and operational effectiveness, achieving results neither could accomplish alone.





Conclusion

Stronger than ever, Trust and Safety stands as the cornerstone of a secure online environment, safeguarding users and providers more efficiently and resiliently. The field is transforming from reactive measures to proactive, AI-driven strategies fortified by data annotation and cutting-edge analytics.

What is crucial is that humans remain an irreplaceable component, offering something even the most sophisticated tools cannot provide—emotional intelligence, nuanced understanding, **humour recognition**, and contextual sensitivity. Together with technology, they create a win-win dynamic where scalability, accuracy, and cultural recognition converge to deliver safer, more inclusive digital experiences.

A FEW WORDS ABOUT CONECTYS

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



Customer Experience

We have 20+ years of experience building global outsourcing CX client projects. We cover all stages, from strategising to setting up digital-first omnichannel contact centres in any location worldwide.



Content Moderation

We provide real-time multilingual moderation services for content, images, and live-streaming videos across social networks, blogs, apps, forums, chat rooms, proprietary customer channels, and beyond.



Data Annotation

We offer a 360° data annotation and labelling solution that ensures precision, scalability, and efficiency throughout the client data journey, empowering the development of ethical, high-performing AI systems.

Our Assets



Over **4000 passionate professionals** working from **14 locations** worldwide, as well as EFH agents.



Excellent services in **35+ languages** available 24/7.



Deep expertise in social media, gaming, retail, e-commerce, fintech, travel and hospitality, startups, software, streaming and more domains.



Contact Us



HQ Address

One Cotroceni Business
Park, Progresului 1,
București 050691,
Romania



US - 1469 532 0215

UK - 44 203 318 1593

EU - 32 929 8011

HK - 852 800 930 130



[sales\(@\)conectys.com](mailto:sales(@)conectys.com)

[info\(@\)conectys.com](mailto:info(@)conectys.com)

**conectys**

www.conectys.com