ActiveFence

The Future is Here:

The State of Trust & Safety 2024

Contents.

Introduction.

For Trust & Safety, 2023 was marked by the rise of generative AI, the passing of key online safety regulations, increased public interest in Trust & Safety issues, and the continued downsizing of many teams.

These shifts have set the stage for the key Trust & Safety themes we expect to see in 2024:

- Democratization of online harms - It is now easier to create, distribute, and disguise harmful content using generative AI.
- Public scrutiny and judicial and regulatory activity around Trust & Safety topics worldwide will drive many strategic decisions for Trust & Safety teams.
- New Trust & Safety companies will continue to emerge and funding activity will grow, supporting the major shift of Trust & Safety teams from “build” to “buy.”

These themes coincide with significant global events and technological advancements. Improvements in generative AI technologies will accelerate the democratization of online harms, amplifying the importance of robust Trust & Safety measures. Meanwhile, major geopolitical events like elections and broad regional conflicts will intensify public and legal scrutiny of Trust & Safety practices. Faced with these challenges, Trust & Safety teams will seek partners to navigate this complex landscape.

We at ActiveFence have made it our mission to be that partner. So, over the past year, we too have adapted, focusing on maximizing impact and efficiency.

By launching and expanding our Trust & Safety platform, **ActiveOS**, and our automated, AI-driven detection system, **ActiveScore**, we are now able to better serve a wider range of companies, allowing organizations large and small to detect and act on more harmful content, faster and easier than ever before. **Our acquisition of SpectrumLabs and reWire** enabled us to further expand our models and support automated detection at scale. The SpectrumLabs acquisition also allowed us to strengthen the broader Trust & Safety community through the **#TSCollective**. Additionally, we unleashed a new **generative AI offering**, numerous **DSA compliance** features, and a **partnership with Agora**, which brings ActiveFence’s offering to real-time video and shopping applications.

Looking ahead to 2024, we know that these changes, and the ones just over the horizon, position us as a true partner for Trust & Safety teams. Read on to gain insights from our team of elite researchers and analysts, at the forefront of online safety, and better prepare yourself for the year ahead.

I hope that you’ll find this a useful resource, and please don’t hesitate to reach out with any questions.

Noam Schwartz
CEO and Co-Founder
ActiveFence

The Rise of Generative AI.

In 2023, generative AI (or GenAI) emerged as a groundbreaking technology that is revolutionizing the digital world.

But alongside its advancements, it has also enabled bad actors to create and spread more advanced forms of harmful content, at scale.

Worryingly, threat actors are using GenAI to disseminate misleading information, fuel extremism, circulate child sexual abuse material (CSAM), and engage in fraudulent activities. These malicious activities involve deepfake creation, intricate phishing tactics, and manipulative social engineering strategies.

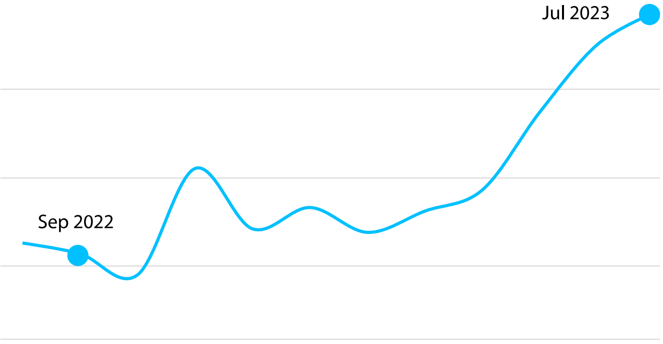
With these advancements in mind, here are the top GenAI-related trends for Trust & Safety in 2024:

1. Exploitation of Multimodal Capabilities:

2024 will see a drastic increase in AI model releases with multimodal capabilities that allow combinations of inputs, including text and images or audio in the same prompt. With these, threat actors can combine two benign prompts to create harmful content.

For example, a non-violative prompt involving adult speech can be combined with childlike audio or imagery. Analyzed separately, these prompts would not raise any flags, but their combination may generate CSAM, which may not be detected using existing models. This threat extends beyond CSAM production, too. Our testing also showed that input combinations can generate personal information of social media users, such as home addresses or phone numbers.

Volume of NCII Detected



212%

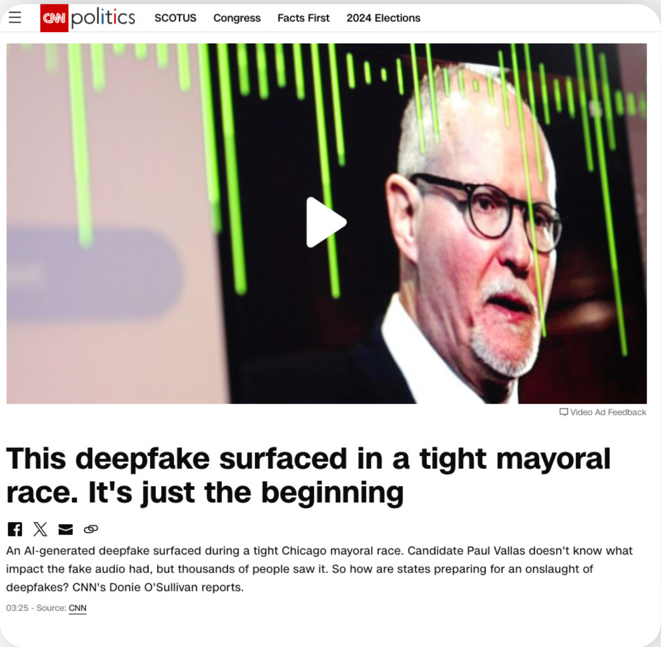
There was a 212% increase in NCII shared on the clear web between 2022 and 2023.

To learn more, read our report.

2. GenAI Audio Impersonation:

Already being used in voice-cloning fraud schemes, text-to-audio models will become more common in 2024.

Extending beyond fraud, this technology may be used by threat actors to clone a prominent politician's voice to falsely claim that they said something controversial or misleading. These types of scenarios can lead to the spread of hate speech and disinformation which in turn could enrage unwitting populations and further widen political divides. Concerningly, ActiveFence has also already documented voice cloning as a tool used by child predators.



This deepfake surfaced in a tight mayoral race. It's just the beginning

An AI-generated deepfake surfaced during a tight Chicago mayoral race. Candidate Paul Vallas doesn't know what impact the fake audio had, but thousands of people saw it. So how are states preparing for an onslaught of deepfakes? CNN's Donie O'Sullivan reports.

03:25 - Source: CNN

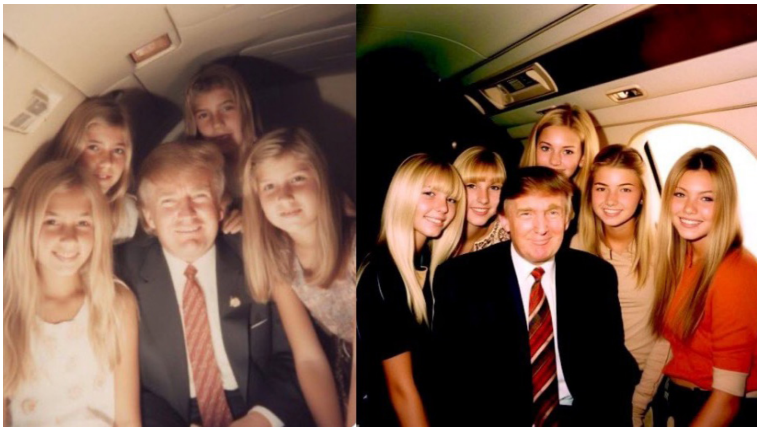
> Former Chicago mayoral candidate Paul Vallas was mischaracterized as a far-right hardliner due to a "hot mic deepfake" that may have cost him the election.
>
> (Source: CNN)

3. Blurred Lines Between Reality and Misinformation:

Since it can create volumes of content, GenAI has an overwhelming capacity to erode the boundaries between fact and fiction. It has now become exceedingly difficult to differentiate between genuine and fabricated content online.

This is especially concerning in an election year where over half of the world will vote. Mis- and disinformation have the potential to fool people into believing false narratives, which can affect election outcomes and how entire populations live day-to-day.

ActiveFence has already reported on the abuse of GenAI to impact elections in the US and the UK, and with GenAI's popularity on the rise, we predict this challenge to grow significantly in 2024.



This picture of former president Donald Trump surrounded by underage girls on Jeffrey Epstein's plane is not real. The AI-generated hoax was shared widely, with one post receiving 1.6M views.

Source: Twitter

GenAI has become a distinct branch of Trust & Safety that demands tailored solutions

Companies that offer GenAI tools for users need to identify and filter risky prompts and outputs, close gaps and loopholes that may cause harm, and tackle intellectual property and copyright-related issues.

At the same time, the use of generative AI has led to an increase in the creation and dissemination of harmful, misleading, and violative user-generated content, making content moderation more difficult and time-consuming for other user-generated content companies.

ActiveFence acknowledges the significant risks posed by GenAI. Drawing from years of Trust & Safety experience, we believe it's possible to mitigate these risks.

Click here to learn more about ActiveFence's AI safety solution.

Challenging Election Integrity.

As we head into 2024, over 70 national elections are poised to take place worldwide, potentially altering the political landscape for more than half of the global population.

The integrity of these elections hangs in the balance due to a variety of challenges. Escalating political polarization among various opposing groups threatens to undermine democratic processes all around the world, exacerbated by online harassment, mass misinformation campaigns, and divisive AI-generated content.

Here are the main trends in election integrity that Trust & Safety teams must pay attention to in 2024:

1. Election Interference Campaigns:

Bad actors have been meddling with domestic and foreign elections for some time now, typically by disseminating false and misleading information regarding various hot-button issues. One issue that always ignites strong emotions is migration. Because there are currently multiple spots around the world undergoing large-scale population shifts due to wars and other crises, there has been an uptick in fear-based misinformation.

One example involves falsehoods related to an influx of Libyan migrants to the Italian island of Lampedusa in September 2023. This event has been used by far-right activists as evidence of the "Great Replacement Theory" in action. This conspiracy theory refers to a secret "deep state" cabal replacing white Americans and Europeans with non-white people. Such narratives can exacerbate racial tensions, promote social unrest, and cause friction in the run-up to the 2024 elections.

260K

One post supporting the "great replacement" narrative received over 260,000 impressions as of September 2023.

To learn more, read our report.

2. GenAI Makes Misinformation Too Easy:

GenAI adds another layer of complexity to the election integrity mix, as it allows bad actors to distribute propaganda and sway public sentiment with easy-to-acquire, easy-to-use AI tools.

In 2023, ActiveFence found both foreign and domestic actors using GenAI to target American voters with divisive narratives. In December 2023 alone, we reviewed politically charged content related to one AI model, with over 83 million impressions. Trust & Safety teams around the world now face the dual task of tackling not only GenAI imagery, but the false narratives spun around them as well.

79K

This AI-generated image of former President Donald Trump getting arrested was widely circulated, with one post reaching 79,000 shares.

Source: PBS

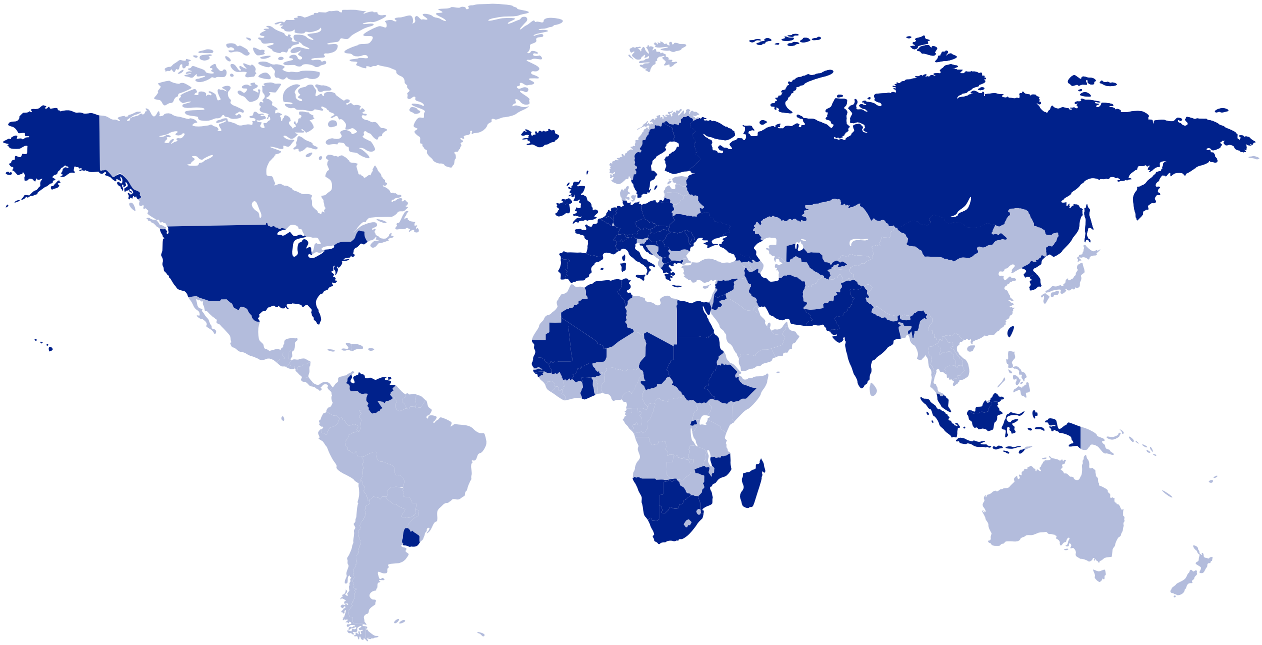
3. Conspiracy Theories and False Claims:

False narratives related to election integrity and alleged election fraud continue to spread widely among users. The upcoming election in the US, given its global influence, merits special attention as divisions continue to intensify around the so-called "culture war", which involves LGBTQ+, abortion, immigration issues, and other topics. All of these topics evoke strong emotions, and conspiracy theories and false claims have been used to fuel and mobilize all sides.

For example, there’s the claim that the Biden Administration has “weaponized the government” by ordering the FBI to intimidate Donald Trump's supporters so they’ll vote for another candidate. There’s also the false narrative surrounding the LGBTQ+ community’s “grooming” of children. These types of false stories further divide populations and erode public confidence in the electoral process.

The stakes are at their highest this year. As we navigate these tumultuous waters, the need for proactive and robust solutions to preserve election integrity has never been more critical. Trust & Safety teams must rise to the challenge of ensuring that fair and transparent elections can take place, free from the disruptive influence of misinformation and disinformation assisted by generative AI.

> Talk to our experts to learn more about how to secure election integrity on your platform.



The results of 2024's elections will impact nearly half of the world’s population.

2023 Saw a Global Rise in Extremism.

In 2023, global events fueled a rise in extremism. Particularly, the escalating Middle East conflict saw a re-emergence of terrorist groups like ISIS and Al-Qaeda, and the surge of hate groups and far-right parties across Europe.

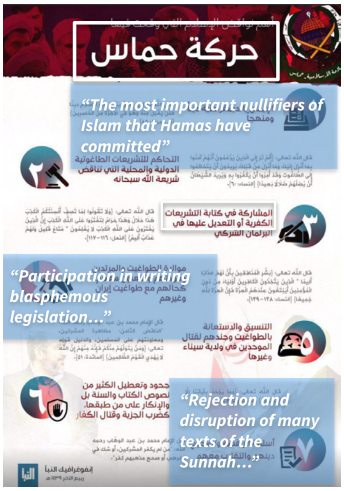
The October 7th attack on Israel sparked more conflict in the Middle East. Yemen's Houthi movement targeted shipping vessels in the Red Sea, while Iranian proxies attacked US and NATO bases. Additionally, ISIS targeted Iran on January 3rd, resulting in numerous casualties.

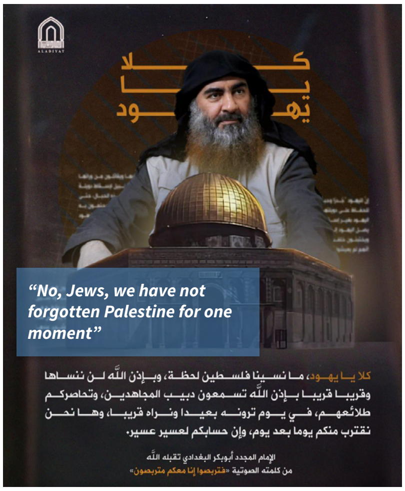
These events also led to an increase in hate speech, propaganda, and misinformation. Trust & Safety teams had a challengine time trying to adapt policies, detection methods, and enforcement to address evolving threats in multiple languages.

These are the trends related to global extremism to look out for in 2024:

1. Al-Qaeda and ISIS are Escalating Online:

Social media platforms have long been hotbeds for extremist activity. We have seen groups like Al-Qaeda and ISIS return to boost their recruitment and propaganda efforts—capitalizing on a downsized Trust & Safety ecosystem and other geopolitical events, like the attack on Israel. T&S teams must learn to identify obscure content related to these groups.


حركة حماس
"The most important nullifiers of Islam that Hamas have committed"
"Participation in writing blasphemous legislation..."
"Rejection and disruption of many texts of the Sunnah..."



140%

Al-Qaeda has increased its propaganda output by 140%.

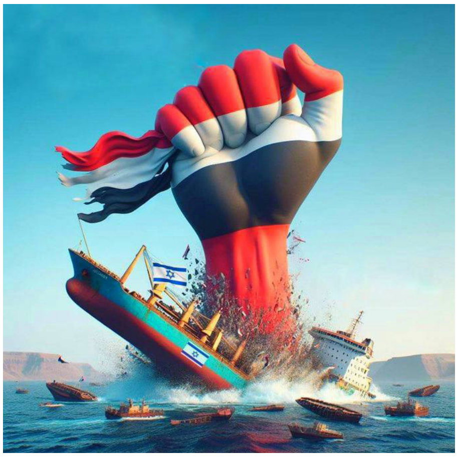
To learn more, read our report.

2. Social Media Continues to Push Dangerous Ideology:

Another startling development in the wake of the Israel-Hamas war was the surge in support for Hamas, which is recognized as a terrorist group by the US and EU. Astonishingly, this increase in support has extended to progressive circles on US college campuses, drawing parallels to the TikTok trend in November 2023 of supporting Osama Bin Laden's “Letter to America.” Written in 2002, the dead Al-Qaeda leader’s antisemitic manifesto reappeared and sparked debate between TikTok users over the Israel-Hamas War. Content like this has led to the legitimization of violence by terror groups and amplified the pressing need for policy-level changes to community guidelines in apps such as TikTok.

3. Extremist Organizations are Leveraging GenAI and Social Media:

Extremist groups like Hezbollah and Hamas have employed AI tools for video production alongside adaptive distribution strategies to broaden their agendas. Hamas in particular used text-to-audio prompts to create various “commentary” videos related to the group’s actions and movements. AI tools help produce these videos with little effort. To distribute them, Hamas has taken advantage of social media platforms—if their accounts are blocked or deleted, they almost immediately set up alternate accounts, or use existing backup accounts. They can also link to their Telegram channel APKs to circumvent any blocks by online app stores. This adaptiveness—combining the speedy creation of content via GenAI, and simple account reactivation after social media blocks—has made moderating content on social apps a game of Whac-A-Mole.



> **Extremists use AI to create propaganda like this image of a Yemini fist crashing through an Israeli ship, which received over 11K views since late last year.**
>
> For more details, read our report.

All these examples paint a chilling picture of the rise of extremist propaganda in late 2023. As we look ahead, 2024 may see these extremist causes and organizations continue to evolve. Documenting these attacks, online social trends, and attempts at manipulating social media to win public support will remain especially crucial for Trust & Safety teams.

Implementing New Regulations.

2023 was a landmark year for online safety regulations, and many of the laws that passed last year are coming into effect in 2024. Trust & Safety teams need to demonstrate willingness to adhere to them, as failure to do so could result in large fines, increased operating costs, and a worse user experience.

That said, the Trust & Safety industry often struggles when it comes to new regulations since the actual requirements are often vague, incomplete, and open to interpretation. Let's look at a few examples:

The UK's Online Safety Act (OSA)
A significant and crucial improvement to online safety legislation, this act outlines several specific requirements and expectations for online platforms, primarily promoting transparency and accountability. However, while the OSA is now in effect, the actual process of fleshing out a more specific code of conduct is ongoing, with a timeline scheduled up to the fourth quarter of 2025. Currently, the OSA could be compared to a menu of potential requirements rather than an absolute prescription for every operator in the UK market.

> For more details on the Online Safety Act, please read The UK's Online Safety Act: What Trust & Safety Teams Can Expect.

India's Digital India Act (DIA)
Looking ahead, upcoming regulations like India's Digital India Act (which is slated for 2025, and will likely be impacted by India's elections this year) will further shape—and complicate—the Trust & Safety landscape. Meant to boost the Indian economy by encouraging more startups to innovate, the DIA also hopes to address the growing rate and sophistication of cyber attacks by providing a more comprehensive and up-to-date framework surrounding online safety, trust and accountability, and the regulation of AI technologies.

The EU's Digital Services Act (DSA):
Designed to unify the European Union under one set of regulations, the DSA is the first set of laws to mandate Trust & Safety as a requirement, not just a suggestion. Under its provisions, online platforms must ban illegal and misleading content, implement a process of content moderation, ensure transparency related to moderation actions, and set guidelines surrounding appeals and flagging of content, among other mandates. It also offers a ray of hope when it comes to regulatory guidance, by providing a database where groups operating in the EU can report the reasons behind content moderation decisions in near-real-time. It is publicly accessible and machine-readable, and updated through a webform or API. This database should lend more clarity to the current landscape, and help pave the way for safer and more accountable online environments.

> For more on DSA compliance, please watch our webinar Unlocking DSA Compliance With 3 ActiveOS Features.

Given the ever-evolving nature of compliance and regulation, the Trust & Safety industry faces ongoing challenges as it works to stay up to date with new laws and requirements. Plus, the inherent global nature of the internet provides even more moving targets when striving to understand and comply with regulations across various jurisdictions. Adapting to these changes while maintaining user trust is a struggle that's only going to intensify in 2024.

The Future Outlook of Trust & Safety.

As 2024 unfolds, the Trust & Safety industry faces a variety of challenges that demand swift action and innovative solutions.

GenAI: One of the most pressing issues in 2024 is the rapid emergence of generative AI tools, which can create increasingly realistic fake content. This poses a significant threat to the safety of online spaces, as identifying and mitigating the spread of manipulated or false information becomes more complex.

Extremism: The battle against extremism remains a top priority, too. Online platforms must continue to enhance their efforts in proactively detecting and removing harmful content to prevent the spread of harmful ideologies.

Elections: Nearly half the world's population will be voting this year. Preserving the integrity of elections is crucial to upholding democratic values. The Trust & Safety industry faces the immense task of combating disinformation campaigns, foreign interference, and AI-generated deepfakes that can undermine the democratic process and erode public trust.

Regulations: The implementation of new online safety regulations in 2024 presents a slew of complications for organizations. Significant challenges that the industry must address include: navigating the requirements, ensuring compliance, and reducing the potential impact on user experience and operational costs.

As such, the Trust and Safety industry faces a tough year in 2024. From combating the risks posed by generative AI and extremism to protecting fragile democracies from misinformation campaigns, these challenges demand innovative approaches, close collaboration between industry stakeholders, and the successful translation of regulatory intent into action.

To stay on top of this year's trends, and to learn more about countering 2024's emerging challenges

Contact us

ActiveFence

Supporting the Future of Trust & Safety

Users thrive in a safe online world.

ActiveFence is the leader in providing Trust & Safety solutions to protect online platforms and their users from malicious behavior and content. Our mission is to protect your platform and users from the broadest spectrum of online harms, unwanted content, and malicious behavior - providing you with the comprehensive data and tools needed to mitigate threats.

The complete Trust & Safety solution

- ✓ **Deep Threat Intelligence:** Proactively avoid risk by accessing expert insights and investigations from the clear, deep and dark web.
- ✓ **ActiveScore:** Accurately detect harmful content and take confident actions, using automated, contextual AI that is fueled by intelligence insights.
- ✓ **ActiveOS:** Simplify moderation orchestration & management using our fully customizable moderation platform, built for scale and efficiency.

Detecting malicious activities across a range of evolving threats

Child Safety • Copyright Infringement • Cyber Crime • Disinformation & Misinformation
Fraud & Scams • Hate Speech • Human Exploitation & Trafficking • Illegal Goods
NCII • Nudity • Profanity • PII • Self Harm • Spam • Terrorism & Extremism • Violence

PROTECTING	ANALYZING	MONITORING	COVERING
3B+	750M+	10M+	100+
USERS	DAILY SIGNALS	SOURCES	LANGUAGES

Why ActiveFence

One place for every Trust & Safety need

From scalable detection to moderation management to proactive investigations, ActiveFence merges **unparallelled human expertise and state of the art technology** to provide Trust & Safety teams with the tools they need to ensure user safety.

Innovative solutions powered by years of expertise

Our diverse team of intelligence and technical experts ensures that every ActiveFence solution is best-in-class.

Intelligence-Led Insights

- Access to **10M sources** on the clear, deep, and dark web
- Deep understanding of evolving threat actor TTPs & the adversarial mindset

AI-Enabled Scale

- Analyzing **750M+ daily signals**
- Our teams of developers and data scientists apply machine learning techniques and mobilize automation

Subject-Matter Expertise

- Academics and experts from **15 abuse areas**
- Elite military-trained OSINT analysts and security researchers
- Technical teams and Trust & Safety veterans

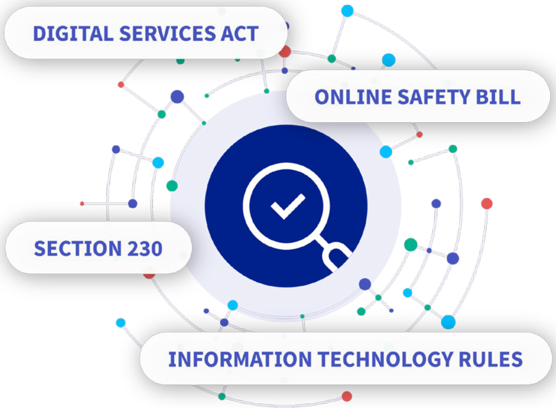
Socio-Political Context

- Linguists & cultural experts across the globe
- Monitoring harm in **100+ languages**
- Tracking geopolitical developments

Ensuring platform compliance

The regulatory landscape is rapidly evolving. Trust & Safety space is in the spotlight as teams face increased legal scrutiny across the world.

ActiveFence provides platforms with dedicated tools for staying compliant while maintaining business continuity and reducing costs.



ActiveFence

About ActiveFence.

ActiveFence is the leading solution for Trust and Safety intelligence and management, protecting online platforms and their users from malicious behavior and content. Trust and Safety teams of all sizes rely on ActiveFence to keep their users safe from the widest spectrum of online harms, unwanted content, and malicious behavior, including child safety and exploitation, disinformation, hate speech, terror, nudity, fraud, and more. We offer a full stack of capabilities with our deep intelligence research, AI-driven harmful content detection, and content moderation platform. Protecting over three billion users globally everyday in 100 languages, ActiveFence lets people interact and thrive online. Backed by leading Silicon Valley investors such as CRV and Norwest, ActiveFence has raised $100M to date, and employs over 300 people worldwide.

activefence.com